المعهد الملكي للثقافة الأمازيغية
ⵉⵏⵙⵜⵉⵜⵓⵜ ⴰⵖⵔⵓⵎⴰⵏ ⵏ ⵜⵙⴷⵓⵔⵉⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Actes de conférence

5ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵙⵓⵔⵉⵜⵜⵉⵏ ⵏ ⵜⵎⵔⵙⵓⵍⵜ ⴷ ⵜⵎⵙⴰⵡⴰⵍ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination

ATAA ALLAH Fadoua

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵜⴽⵏⵓⵍⵓⵊⵢⵜ ⵏ ⵓⵙⵎⵔⵙ ⴷ ⵓⵎⵙⵙⵓⵍ

الأمـازيغية وتكنولوجيـا المعلـومـات والتـواصـل

Technologies d'Information et de Communication pour l'Amazighe

المعهد الملكي للثقافة الأمازيغية
ⵜⴰⵙⴷⴰⵡⵉⵜ ⴰⵖⵔⵉⴼⵏ ⵏ ⵜⵣⵔⴼⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Actes de conférence

5ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵉⵎⵓⵏⵉⵜⵜ ⵏ ⵜⵙⵏⵜⵔⵜ ⵏ ⵜⵎⵙⵙⵓⵔⵉⵜ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination
Ataa Allah Fadoua

Publications de l'Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Série : Colloques et séminaires N° 40

Titre *: Technologies d'Information et de Communication pour l'Amazighe*
Coordination *: Ataa Allah Fadoua*
Conception *: Nadia Kiddi (Unité de l'édition)*
Editeur *: Institut Royal de la Culture Amazighe*
Imprimerie *: El Maârif Al Jadida - Rabat*
Dépôt légal *: 2015MO1163*
ISBN *: 978-9954-28-184-0*
ISSN *: 2421-9711*
Copyright *: ©IRCAM*

PREFACE

Le présent ouvrage constitue un recueil des actes de la 5ème édition de la conférence internationale sur les technologies d'information et de communication pour l'amazighe, qui s'est tenue à Rabat les 26 et 27 novembre 2012, au siège de l'Institut Royal de la Culture AMazighe (IRCAM). Il expose le fruit d'une rencontre entre chercheurs et jeunes chercheurs de disciplines variées qui ont entrepris de s'adresser à tous ceux qui s'interrogent autour de l'informatisation et le traitement automatique des langues peu dotées, et en particulier la langue amazighe.

Les contributions ici transcrites se révèlent à la fois divergentes et complémentaires. Elles permettent de mettre en évidence le caractère multidisciplinaire des travaux, la richesse des aspects méthodologiques et la diversité des différents aspects théoriques et pratiques. Ces contributions représentent quatre vingt trois pour cent des articles soumis à cette conférence, et qui ont fait l'objet d'une évaluation par au moins deux chercheurs experts du sujet traité. Elles sont structurées selon cinq grands thèmes à savoir : Outils de traitement automatique des langues, Recherche d'information et extraction des connaissances, Reconnaissance optique des caractères, Ressources linguistiques, et Traitement de la parole.

Nous tenons à remercier chaleureusement tous les intervenants de cette conférence, venus d'horizons divers, qui ont généreusement partagé avec nous leurs travaux de recherche. Nous souhaitons aussi témoigner notre reconnaissance à tous les membres du Comité de Lecture qui ont cru en cet événement, soutenu ces travaux et présidé les sessions de ces deux journées scientifiques. J'en profite également pour remercier toute l'équipe du Comité d'organisation qui s'est impliquée dans la préparation et la réussite de cette conférence. Que soient, ici aussi, remerciés tous ceux qui ont contribué, de près ou de loin, à ce que cet événement prenne de l'ampleur tout au long de ses années d'existence.

Nous espérons que ces actes de conférence deviendront pour tous chercheurs et experts une référence sur les travaux de recherche et les études réalisés pour l'amazighe dans le domaine des technologies de l'information et de la communication, en vue d'enrichissement de la réflexion et de la production dans ce domaine.

Comité de lecture

Aboutajdine Driss (FSR, Rabat)
Aoughlis Farida (UMMTO, Tizi Ouzou)
Bennani Samir (EMI, Rabat)
Bouyakhf Houssaine (FSR, Rabat)
Cavalli Sforza Violetta (AUI, Ifrane)
Cherkaoui Chihab Dine (ENCG, Agadir)
El Hamdani Abdelfettah (IERA, Rabat)
El Jihad Abdelhamid (IERA, Rabat)
El Mouradi Abdelhak (ENSIAS, Rabat)
Fadouli Nouredine (EMI, Rabat)
Fakir Mohamed (FST, Beni Mellal)
Jean Thierry (OLPC, France)
Idrissi Najlae (FST, Beni Mellal)
Khalidi Idrissi Mohamed (EMI, Rabat)
Mammas Driss (EST, Agadir)
Modi Issouf (BEM, Nigeria)
Pognan Patrice (INALCO, Paris)
Rachidi Ali (ENCG, Agadir)
Rami Salim (FPS, Safi)
Rosso Paolo (UPV, Valence)
Semmar Nasredine (CEA, Paris)
Soudi Abdellah (ENIM, Rabat)
Tigziri Noura (UMMTO, Tizi Ouzou)
Yousfi Abdellah (FS, Rabat)
Zenkouar Lahbib (EMI, Rabat)

Comité d'organisation

Ait ouguengay Youssef
Ataa Allah Fadoua
Boulaknadel Siham
Boumediane Mounia
El Hamdaoui Amal
El Marssi Karim
Frain Jamal

Table des matières

Vue Générale sur les Technologies du Langage Humain: Cas des Projets Européens NEMLAR & MEDAR **11**
Abdelhak MOURADI

Base de Données Numérique de la Terminologie de Spécialité en Tamazight **21**
Noura TIGZIRI, Ramdane BOUKHERROUF

Les Humanités Numériques et le Projet de Bibliothèque Numérique Franco-Berbère **27**
Alain VAUCELLE, Tassadit YACINE

Problématiques d'Usage et d'Intégration des Langues Peu Dotées dans le Web des Données Ouvertes (Linked Open Data ou LOD) : Cas de l'Amazighe **39**
Hammou FADILI

Vers un Dictionnaire Electronique de l'Amazighe **55**
Fatima Zahra NEJME, Siham BOULAKNADEL, Driss ABOUTAJDINE

Initiative pour le Développement d'un Corpus de la Langue Amazighe **65**
Siham BOULAKNADEL, Fadoua ATAA ALLAH

De la Taxinomie au Traitement Automatique des Verbes en Amazighe **75**
L'houssaine EL GHOLB

Projet GCAM : Vers une Gestion Informatisée du Corpus Amazighe à l'IRCAM **89**
Youssef AIT OUGUENGAY, Amal EL HAMDAOUI, Brahim El HASNAOUY, Abdellah FADDAH

Etude Statistique sur les Erreurs d'Edition dans la Langue Arabe **97**
Hicham GUEDDAH, Abdellah YOUSFI

Analyseur Morphologique des Mots Arabe en Utilisant le Dérivé et Schème de Surface **103**
Said IAZZI, Abdellah YOUSFI, Mostafa BELLAFKIH

Automatic Tamazight Spelling Correction Using Noisy Channel Model and Bigram Language Model **113**
Said GOUNANE, Mohamed FAKIR, Belaid BOUIKHALEN

Réalisation d'un Editeur de Texte pour la Langue Amazighe et d'une Barre d'Outils Amazighe pour MS Office 119
Nourdine AIT ZENGUI, Ali RACHIDI, El Houssine BOUYAKHF

ANMorph : Amazigh Nouns Morphological Analyzer 127
Hanae RAISS, Violetta CAVALLI-SFORZA

Caractérisation de Voyelles et d'Emphatiques Berbères en vue d'une Identification de Locuteurs Dépendant de Texte 137
Fatma zohra CHELALI, Amar DJERADI, Hocine TEFFAHI

Prosodie des Phrases Interrogatives et Affirmatives en Langue Berbère 147
Ramzi HALIMOUCHE, Hocine TEFFAHI, Leila FALEK

Effect of Dialect, Size of Population and PCA on Speaker Verification Performance 157
Kawthar Yasmine ZERGAT, Abderrahmane AMROUCHE, Nassim ASBAI

La Transcription Phonétique de la Langue Amazighe : Vers un Système de Synthèse de la Parole 165
Hassan JAA, Belhaj ELGRAINI

Reformulation des Requêtes pour le Domaine de l'Enseignement : Cas des Cours de l'Algorithmique 173
Mohamed RACHDI, El Habib BEN LAHMAR, El Hassan LABRIJI, Aziz CHARA, Kamel ET GUEMMAT

Recherche par Expression dans le Texte Intégral Multilingue 183
Yahya HLAL

Ouverture des Noms de Domaine Internet Génériques par l'ICANN (New Generic Top Level Domains) 193
Ali BOUALLOU

Système de Recherche d'un Mot-Image : Cas des Mots Arabes Imprimés 207
Anass SMAILI, Ali LASFAR, Mohamed SBIHI

Enrichissement Sémantique des Requêtes Multi-Mots 215
Mohamed RACHDI, El Habib BENLAHMAR, El Hassan LABRIJI

Matching des Documents XML par la Mesure de Similarité à Base WordNet 225
Fatiha DJAHAFI, Abdelkader HAOUAS

Cours de Tamaziγt par Internet : Problèmes et Propositions 233
Saïd CHEMAKH, Malika SABRI

L'Usage de la Langue Amazighe dans les Médias Algériens 237
Ouerdya KIRECHE

Contribution à la Reconnaissance des Documents Tifinaghes 241
Mehdi BOUTAOUNTE, Mohamed FAKIR, Belaid BOUIKHALENE

Hybridation des Modèles de Markov Cachés et de la Logique Floue pour la Reconnaissance des Caractères Tifinaghes Manuscrits 251
Aissa HAIDAR, Mohamed FAKIR, Omar BENCHAREF

Recognition of Amazigh Characters Using Visual Rotation Invariant Features 261
Younès RAOUI, El Houssine BOUYAKHF

Réalisation d'un OCR pour l'Ecriture Amazighe Imprimée 269
Youssef ES-SAADY, Mustapha AMROUCH, Ali RACHIDI

Study of Effect of Regularized Neural Network on the Accuracy of Handwriting Recognition 279
Meena M. MAKARY, Inas A. YASSINE

Nouvelle Méthode de Reconnaissance Automatique de Caractère Tifinaghe 289
Mustapha AMROUCH, Youssef ES-SAADY, Ali RACHIDI, Mohamed EL HAJJI, Mostafa EL YASSA, Driss MAMMASS

Vue Générale sur les Technologies du Langage Humain
Cas des Projets Européens NEMLAR & MEDAR

Abdelhak Mouradi
Université Mohammed V Souissi, ENSIAS
mouradi@ensias.ma

Résumé

Dans cet article, nous allons donner un aperçu sur les technologies du langage humain en adressant les deux composantes : le langage écrit et le langage parlé. Nous focaliserons notre attention sur les ressources linguistiques et les outils de traitement du langage pour la langue arabe. Nous exposerons également les travaux menés dans le cadre de deux projets européens auxquels nous avons participé, à savoir : NEMLAR et MEDAR.

1. Introduction

L'être humain a toujours aspiré à pouvoir communiquer avec la machine en utilisant le moyen le plus naturel et convivial à savoir la parole et à apprendre aux ordinateurs à traiter le langage écrit comme il est capable de le faire.

Plusieurs laboratoires de recherche se sont intéressés à ce domaine et une nouvelle discipline a vu le jour, à savoir le « Traitement Automatique du Langage Naturel » TALN (ou NLP : Natural Language Processing) qui fait appel à l'informatique, à la linguistique et à l'intelligence artificielle.

Les technologies permettant de répondre à ces problématiques et de réaliser ces applications sont appelées « Technologies du Langage Humain » (HLT / Human Language Technologies).

2. Les technologies de traitement automatique de la parole

Ces technologies connaissent un grand succès. Plusieurs systèmes sont aujourd'hui capables de produire de la parole de bonne qualité tout en ayant un naturel acceptable par l'être humain. Des systèmes de synthèse de la parole à partir du texte (Text To Speech Systems) dans différentes langues sont commercialisés. La liste des applications des systèmes de synthèse est assez fournie mais il reste encore un effort à faire au niveau de la prosodie pour améliorer le naturel de la parole produite.

D'un autre côté, les systèmes de reconnaissance de la parole connaissent un grand essor et ont transité par plusieurs phases. Ainsi, on est passé des systèmes de reconnaissance de quelques mots isolés pour un locuteur donné à des systèmes de reconnaissance de la parole continue, multilocuteurs et à vocabulaire très large. Parmi les applications phares de la reconnaissance de la parole on peut citer la « dictée vocale » et « la transcription d'enregistrements sonores ».

Aujourd'hui, la majorité des systèmes de reconnaissance de la parole est basée sur l'approche statistique. Ces systèmes sont généralement constitués de deux unités principales, le module de décodage acoustico-phonétique et le module de modélisation du langage.

3. Les technologies de traitement automatique des langues

Ces technologies permettent de gérer et d'exploiter les documents textuels disponibles sous une forme électronique.

Elles mettent en jeu plusieurs niveaux de l'analyse du texte : morphologique, lexicale, syntaxique, sémantique et pragmatique. Elles ont donné lieu au développement de plusieurs applications couvrant un large spectre allant de la veille technologique à la traduction automatique des langues.

Les systèmes de traitement automatique des langues s'appuient sur plusieurs types de ressources linguistiques et d'outils informatiques permettant l'acquisition, la préparation, la collecte, la gestion et l'utilisation de ces ressources. Parmi ces outils et ressources nous pouvons citer :

- Les corpus oraux et écrits annotés ou non ;
- Les dictionnaires ;
- Les analyseurs morphologiques ;
- Les correcteurs orthographiques ;
- Les analyseurs syntaxiques ;
- Les correcteurs grammaticaux ;
- Les systèmes de traduction automatique ;
- Les systèmes de reconnaissance de la parole ;
- Les systèmes de synthèse vocale ;
- Les moteurs de recherche.

Les systèmes de transcription automatique d'informations radio et télédiffusées.

4. Les projets européens NEMLAR & MEDAR

La production des ressources et des outils linguistiques est primordiale non seulement pour le développement économique des pays mais également pour leur enrichissement culturel.

En se focalisant sur la technologie d'une langue (arabe, amazighe, …) et en rendant disponibles les outils et les contenus nécessaires à son traitement automatique, son utilisation va croitre et sa dépendance vis-à-vis des autres langues va baisser.

C'est dans ce cadre qu'un consortium s'est constitué pour le développement de la langue arabe, consortium qui a bénéficié du soutien financier de la commission européenne.

Ainsi, deux projets ont été menés, à savoir :

NEMLAR : Network for Euro-Mediterranean LAnguage Resources

MEDAR : MEDiterranean ARabic language and speech technology

5. 1. Le projet NEMLAR

NEMLAR est un projet soutenu par la Commission Européenne, avec des partenaires de l'UE et des pays parlant la langue arabe dans la région méditerranéenne.

5.1.1. Objectifs du projet

Le projet s'est fixé comme objectifs de :

- Dresser l'état de l'art des ressources linguistiques et d'outils pour l'arabe dans la région ;
- Elaborer une définition BLARK (Basic LAnguage Resource Kit) pour l'arabe ;
- Commencer le développement de ressources linguistiques ou la mise à jour des ressources linguistiques existantes ;
- Créer de la visibilité pour la technologie de la langue arabe, à travers un bulletin d'information et par le biais d'une conférence internationale ;
- Constituer un réseau de partenaires européens et méditerranéens pour étudier, évaluer, échanger les meilleures pratiques et pour stimuler le développement adéquat de ressources linguistiques en arabe et dans autres langues autochtones dans la région méditerranéenne, afin d'assurer l'interopérabilité entre les langues des pays européens et méditerranéens.

5.1.2. Détails sur le projet

Date de début : 01 03 2003
Date de fin : 31 07 2005
Durée : 30 mois
Coût du Projet : 500.000 EURO
Type du Programme : Cinquième programme cadre (Union Européenne)
Type du contrat : Mesures d'accompagnement et de support

Coordinateur : University of Copenhagen Danmark

Organismes participants :

The Open University United Kingdom
Centre National de la Recherche Scientifique Lyon France
Utrecht University Nederland
Institute for Language and Speech Processing Athens Greece
Amman University Jordan
European Language Resources Distribution Agency Paris France
Université Lumière Lyon 2 France
IT.COM, Information Technology Communications Tunis
ENSIAS - Universite Mohammed V Souissi Rabat
The Engineering Company for Computer Systems Development Egypt
University of Balamand Tripoli Lebanon
Birzeit University Birzeit West Bank and Gaza Strip

5.1.3. Résultats du projet

Nous décrivons ci-dessous les résultats les plus marquants du projet NEMLAR.

◆ Kit de ressources linguistiques de base BLARK

Dans le cadre du projet NEMLAR, une première étude a été menée pour recenser les ressources linguistiques disponibles en arabe. Une deuxième étude a eu pour objet de déterminer les besoins des industriels en ressources linguistiques dans la région de la méditerranée.

D'autre part, le premier BLARK arabe, qui décrit l'ensemble minimal de ressources linguistiques nécessaires pour le développement des technologies du langage arabe, a été élaboré. Il est essentiellement destiné aux chercheurs dans les universités, à l'industrie et aux formateurs.

En se basant sur le BLARK arabe et sur les ressources linguistiques disponibles, on peut déterminer les ressources qui manquent et établir une liste des priorités des composantes à produire.

◆ Ressources linguistiques développées

Plusieurs ressources linguistiques ont été développées dans le cadre du projet et notamment :

■ Un corpus oral d'actualités radiophoniques NEMLAR

Le corpus oral d'actualités radiophoniques NEMLAR est composé d'environ 40 heures d'émissions radiophoniques en arabe standard. Les émissions ont été enregistrées depuis quatre stations de radio différentes : Medi1, Radio Orient, RMC – Radio Monte Carlo, RTM – Radio Télévision Maroc. Chaque transmission contient entre 25 et 30 minutes d'actualités et d'interviews. Les enregistrements ont été effectués à trois périodes différentes entre le 30 juin 2002 et le 18 juillet 2005. Tous les fichiers ont été enregistrés au format linéaire PCM, 16 kHz, 16 bits.

■ Un corpus écrit NEMLAR

Le corpus écrit NEMLAR est constitué de 500 000 mots de texte arabe regroupés en 13 catégories différentes, visant à obtenir un corpus bien équilibré qui offre une représentation de la variété de traits syntaxiques, sémantiques et pragmatiques de la langue arabe moderne.

Les différentes catégories sont :

- Actualités politiques : 48 000 mots
- Débat politique : 30 000 mots
- Texte Islamique (prières et autres) : 29 000 mots
- Expressions de mots communs : 8 500 mots
- Textes extraits d'émissions radiophoniques : 5 500 mots
- Affaires : 20 000 mots
- Littérature arabe : 30 000 mots
- Actualités générales : 100 000 mots
- Interviews : 56 000 mots

- Presse scientifique : 50 000 mots
- Presse sportive : 50 000 mots
- Explications d'entrées de dictionnaire : 52 000 mots
- Texte du domaine juridique : 21 000 mots

■ Un corpus de synthèse de parole NEMLAR

Le corpus de synthèse de parole NEMLAR comprend les enregistrements de 2 locuteurs ayant l'arabe égyptien comme langue maternelle (homme et femme, respectivement de 35 et 27 ans), réalisés dans un studio depuis 2 canaux (voix et laryngographe). Les enregistrements sont constitués de plus de 10 heures de données avec leurs transcriptions. Les échantillons de parole sont stockés en 96 kHz, 24 bits. Le locuteur a lu 2 032 phrases énoncées, couvrant environ 42 000 mots en trois catégories : parole transcrite (6,600 mots - 20%), texte écrit (16,500 mots - 50%), et phrases construites (10,300 - 30%).

La parole transcrite consiste en du texte de différents domaines. Le texte écrit est composé d'extraits de phrases courtes d'actualités, de romans et d'histoires courtes. Chaque paragraphe est présenté sur une feuille d'énoncé (prompt).

Les phrases construites sont constituées de phrases fréquentes et de phrases pour la couverture de diphones. Les phrases fréquentes sont obtenues à partir de textes écrits (articles, actualités, etc.).

La base de données est fournie avec la transcription orthographique, prosodique et phonétique en SAMPA. Un lexique de prononciation comprenant 3 589 mots avec leur représentation phonétique en SAMPA est également disponible.

5.1.4. Sensibilisation et dissémination des résultats

Le projet s'est intéressé en premier lieu aux différentes formes de sensibilisation et de dissémination en identifiant les connaissances à diffuser, les groupes à cibler et les canaux adéquats.

D'abord, il a été crée un site web ainsi qu'un logo spécifiques au projet (www.nemlar.org).

Ensuite, le projet a été référencé dans les publications des partenaires lorsque cela était approprié.

Dans le cadre de la dissémination des résultats, NEMLAR a tenu sa première conférence sur les ressources et les outils de la langue arabe les 22 et 23 Septembre 2004 au Caire. Elle a réuni des chercheurs des mondes de l'université et de l'industrie de partout dans le monde pour discuter des questions des technologies du langage humain en arabe.

5. 2. Le projet MEDAR

5.2.1. Objectifs du projet

Le consortium du projet MEDAR s'est fixé 4 objectifs primordiaux qui sont :

- La consolidation d'un réseau d'acteurs dans tous les domaines des technologies du langage humain ;

- Le développement d'une feuille de route de la coopération fondée sur une vision claire de l'évolution prévisible du marché, des potentiels technologiques et des possibilités de coopération ;
- La mise à jour du BLARK en s'intéressant en particulier à l'ensemble minimal de ressources et d'outils nécessaires pour effectuer de la recherche et de la formation sur les ressources linguistiques et les technologies du langage humain tout en mettant l'accent sur la traduction automatique (MT : Machine Translation) et la recherche d'information multilingue (MLIR : Multilingual Information Retrieval) ;
- Le soutien au développement d'outils et de ressources, en particulier pour la traduction automatique et la recherche d'information multilingue sur la base des technologies des partenaires et des logiciels libres.

5.2.2. Détails sur le projet

Date de début : 01 02 2008

Date de fin : 31 07 2010

Durée : 30 mois

Coût du projet : 798.553 EURO

Type du programme : Septième programme cadre

Type du contrat : Coordination and support actions

Coordinateur : Université de Copenhague Danemark

Organismes participants :

Evaluations and Language Resources Distribution Agency France
University of Balamand Lebanon
Al-Ahlyya Amman University Jordan
Universiteit Utrecht Pays Bas
Institute for Language and Speech Processing / "Athena" R.C. Greece
The Engineering Company for Digital Systems Developement Egypt
Birzeit University Palestinian-Administered Areas
ENSIAS - Universite Mohammed V-Souissi Morocco
Commissariat à l'Energie Atomique et aux Energies Alternatives France
Centre National de la Recherche Scientifique France
The Open University United Kingdom
Université Lumière Lyon 2 France
Ibm World Trade Corporation Joint Stock Company Egypt
Sakhr Software Co. Egypt

5.2.3. Résultats du projet

Les principaux résultats du projet MEDAR sont :

- Une base de connaissances des acteurs, des produits et des ressources et une mise à jour du BLARK arabe.

Suite aux résultats d'une enquête et ceux recueillis auprès des partenaires à propos des applications existantes relatives aux technologies du langage humain en arabe, une base de connaissances a été construite à partir des données récoltées. Elle contient des informations détaillées et utiles sur les organismes, les produits, les services et les ressources linguistiques.

Elle est accessible à partir du lien http://www.elda.org/medar_knowledge_base/.

◆ Une feuille de route qui a pour objet de décrire les domaines et les priorités de la collaboration entre les pays européens et les pays parlant la langue arabe, ainsi que la coopération, d'une manière générale entre les pays, entre les universités et aussi entre les universités et les industriels.

Cette coopération doit conduire à renforcer la communauté impliquée dans les technologies de la langue arabe, et à mettre à disposition plus de technologies et de produits en arabe.

◆ Un exemple concret de mise en œuvre de la feuille de route de la coopération, pour le domaine de l'enseignement supérieur.

En effet, la formation universitaire constitue un point important pour le développement des technologies avancées pour la langue arabe. Les acteurs dans le domaine doivent disposer de ressources humaines fortement qualifiées. L'accent mis sur l'éducation provient de son importance en tant que force motrice pour le développement des ressources humaines.

La coopération dans le domaine de l'enseignement supérieur peut se faire à deux niveaux :

- Le niveau formation académique, en incorporant des formations au niveau licence, master et doctorat en relation avec les champs d'intérêt des technologies du langage humain et en particulier de la langue arabe.
- Le niveau recherche, en focalisant une partie de cette recherche sur les axes pertinents des technologies de la langue arabe.

Une recherche avancée va contribuer à combler le fossé avec l'industrie du logiciel et à convaincre les industriels à s'impliquer davantage et à développer des produits pour le marché.

Dans un premier temps, quatre institutions ont été identifiées par le consortium pour établir un curriculum commun. Il s'agit de :

- La Faculté des Technologies de l'Information à l'Université d'Amman ;
- L'Académie Arabe des Sciences, des Technologies et du Transport Maritime, Faculté d'Informatique et des Technologies de l'Information, le Caire ;
- L'Université de Balamand, Liban ;
- L'École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V Souissi, Rabat.

◆ Un package pour la traduction automatique

Les objectifs du projet visent la sélection d'un certain nombre d'outils, leur adaptation et leur personnalisation à la langue arabe (en tant que source ou langue cible), leur évaluation et le développement de nouvelles ressources linguistiques pour leur amélioration.

Le travail s'est déroulé en deux phases :

La première avait pour but de développer un système de traduction automatique de base. Après plusieurs études, le choix du consortium s'est porté sur l'adaptation du système de traduction automatique MOSES. Les détails sur ce logiciel peuvent être obtenus à partir du lien http://www.statmt.org/moses/?n=Moses.Background

Parmi les raisons justifiant le choix de MOSES on peut citer :

- C'est un logiciel libre basé sur une approche statistique ;
- Il a une qualité comparable à celle des systèmes propriétaires ;
- Il a été utilisé pour plusieurs autres paires de langages ;
- Il peut contenir facilement d'autres informations linguistiques ;
- Il a les capacités de supporter la traduction automatique parole-parole car il peut prendre en entrée les résultats de la reconnaissance automatique de la parole.

La deuxième phase consiste en l'amélioration du système de base à la fois avec de nouvelles ressources linguistiques (nouveaux corpus parallèles) et avec de nouveaux outils (outils d'alignement de mots des partenaires). Parmi les tâches de cette phase, nous pouvons citer :

- Construire un corpus parallèle anglais ⬅➡ arabe, soit pour enrichir ce qui a été utilisé dans la phase 1 ou bien pour couvrir un nouveau domaine (santé, économie, etc.) ;
- Aligner le corpus en utilisant certains des aligneurs des partenaires en plus de l'aligneur Giza ++ ;
- Collecter de nouveaux corpus en arabe et en anglais afin d'entrainer les modèles de langage ;
- Exécuter les systèmes MOSES et ceux des partenaires avec les nouveaux ensembles de données.

◆ Un package pour l'évaluation de la traduction automatique

Le contexte est d'évaluer les systèmes de traduction de l'anglais vers l'arabe réalisés et d'identifier leurs performances.

Les données sources pour les campagnes d'évaluation de la traduction automatique ont été collectées. Elles proviennent de sites Internet traitant du thème du changement climatique.

Les ressources linguistiques, qui se composent des textes traduits (appelés « textes de référence »), ont été traduites par quatre différents traducteurs professionnels confirmés. Ces textes représentent la cible à laquelle les résultats de traduction automatique seront comparés.

Les métriques d'évaluation BLEU (Bilingual Evaluation Understudy), NIST (National Institute of Standards and Technology) et WER (Word Error Rate) ont été utilisées pour évaluer la sortie du système de traduction automatique vis-à-vis des traductions de référence.

5.2.4. Sensibilisation et dissémination des résultats

La sensibilisation à l'intérêt de développer les technologies de la langue et de la parole pour l'arabe, la diffusion des connaissances acquises à travers les enquêtes et la dissémination des résultats du projet constituent un objectif majeur de ce dernier.

Plusieurs canaux de communication ont été utilisés, notamment :

- Le site web www.medar.info et les sites web des partenaires ;
- Un dépliant décrivant le projet a été largement diffusé ;
- La presse et la télévision (Al Jazeera) ;
- La lettre d'information éditée tous les trois mois ;
- Les rencontres et les forums ;
- La 2ème conférence internationale sur les ressources et outils de langue arabe qui a eu lieu au Caire les 22 et 23 Avril 2009. En marge de cette conférence, trois tutoriels sur les technologies du langage naturel et la traduction automatique ont été organisés ;
- Deux ateliers sur les technologies de la langue arabe ont eu lieu lors des conférences LREC 2008 à Marrakech et LREC 2010 à Malte.

6. Conclusion

Dans cet article nous avons donné, à travers les deux projets NEMLAR et MEDAR financés dans le cadre du programme ICT de la Commission Européenne, une vue synthétique sur les ressources et les outils de base nécessaires au développement des technologies du langage pour l'arabe.

Ce travail a été réalisé grâce à une coopération internationale entre des pays l'Union Européenne et des pays du bassin méditerranéen. Il constitue un jalon pour tous les acteurs dans le domaine des technologies de la langue et de la parole et peut constituer une base et une référence pour d'autres langues naturelles et particulièrement pour la langue amazighe.

Bibliographie

« Building Annotated Written and Spoken Arabic LRs in the NEMLAR project » . Article présenté à LREC, mai 2006.

http://catalog.elra.info

http://www.medar.info/

http://www.statmt.org/moses/?n=Moses.Background

Base de Données Numérique de la Terminologie de Spécialité en Tamazight

Noura Tigziri **Ramdane Boukherrouf**
Département de langue et culture amazigh de Tizi-Ouzou.
Université Mouloud Mammeri de Tizi Ouzou
nora.tigziri@gmail.com boukherouf@yahoo.fr

Résumé

Notre contribution, consiste à présenter une base de données numérique consultable sur internet, de la terminologie de spécialité en Tamazight dans toutes ses variétés. En effet, le projet en question contient toutes les propositions en matière de terminologie de spécialité en tamazight (linguistique, littérature, civilisation, informatique, medias, terminologie scolaire, etc.).

1. Présentation du projet

Notre proposition de communication consiste à présenter une partie d'un projet que nous menons actuellement dans le cadre des projets nationaux de recherche, dirigé par le professeur TIGZIRI Noura, domicilié au laboratoire d'aménagement et d'enseignement de la langue amazighe de département de langue et culture amazighes de Tizi Ouzou.

Le projet en question, consiste à élaborer un dictionnaire électronique (consultable sur internet) des différentes propositions de la langue amazighe dans toutes ses variétés. Ce dictionnaire passera inévitablement par la mise en place d'une base de données de la langue amazighe contenant la terminologie de spécialité (linguistique, littérature, civilisation, informatique, medias, terminologie scolaire, etc.).

La langue amazighe qui était jusqu'à un passé récent, une langue essentiellement orale, a vu son passage l'écrit et son enseignement rencontrer d'énormes problèmes dus essentiellement à un manque d'outils didactiques tels les dictionnaires. S'il est vrai que des glossaires ont vu le jour, que des terminologies foisonnent sur le terrain, il n'en demeure pas moins, qu'aucun de travail de collectes, de dépouillement, d'analyse de toutes ces données n'a été réalisé jusqu'à ce jour afin de disposer d'une source complète.

Les raisons de cet état remontent au statut même de la langue amazighe. Même si elle a été introduite dans le système éducatif en 1995, c'est une langue encore non normalisée et standardisée. Aussi, lors d'une confection d'un dictionnaire pour ne citer que cet élément, le chercheur ne fait pas uniquement un travail de mise en place d'un dictionnaire avec tous les outils théoriques et méthodologiques qui s'imposent, mais il élabore aussi un travail d'aménagement de l'écriture et du lexique.

2. Problématique

Les matériaux utilisés proviennent d'un dépouillement systématique de toutes les sources existantes (glossaires, lexiques, manuels, etc.) et d'enquêtes sur le terrain pour compléter la base de données, notamment la terminologie utilisée par les enseignants au niveau des établissements de l'éducation (primaire, moyen et secondaire) et des départements de langue et culture amazighes(Tizi Ouzou, Bouira et Bejaia).

3. Objectifs du projet

Cette recherche appliquée a plusieurs objectifs :

1- Mettre à la disposition d'apprenants un dictionnaire aussi complet que possible de la langue amazighe, en regroupant toutes les sources existantes mais éparpillées.

2- Connaître la variation et la prendre en charge dans la base de données. Ce qui conduira forcément à un travail d'aménagement, principalement du lexique et de la grammaire.

3- Accompagner l'intégration de l'amazighe dans l'enseignement, la recherche et les médias algériens en mettant à la disposition des formateurs et des usagers de la langue amazighe une base de données amazighes qui répond à leurs besoins.

4. Méthodologie

Cette recherche comprendra deux volets essentiels

1- Elaboration de la base de données électronique elle-même. Cette phase nécessitera l'intervention d'un informaticien et de linguistes berbérisants.

2- La deuxième phase consiste au dépouillement de toutes les sources existantes.

3- La troisième phase consiste à faire des enquêtes sur le terrain.

4- L'alimentation de la base de données.

En plus du lexique, cette base de données contiendra un certain nombre de données qui sont :

- Des données sur les structures formelles et sémantiques qu'engage le lexique.
- Des données géolinguistiques : les tendances régionales en matière lexico-sémantique.
- Des données sur les racines des mots.
- Des données sur les emprunts.
- Des données sur la grammaire.
- Des données de prononciation régionale.
- Des données de morphosyntaxe.

5. Présentation de la base de données

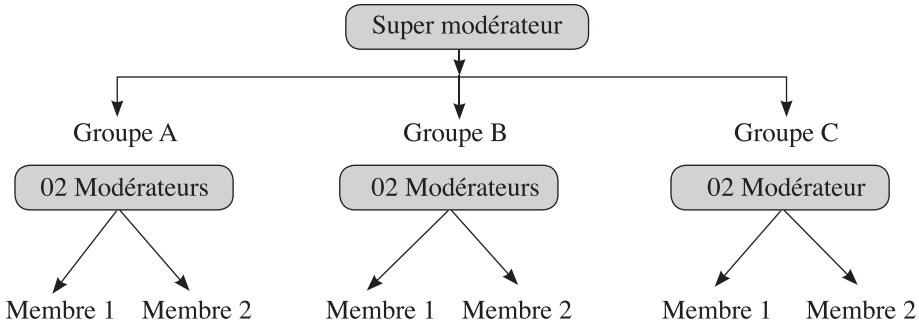
Cette étape consiste à présenter la base de données en décrivant sa structure et toutes les étapes à suivre lors de la saisie des lexies.

5.1. La structure de la base de données

Les agents qui s'occupent de l'alimentation de la base de données sont divisés en trois catégories, et ce, selon leurs statuts dans la base de données.

- Le membre : se charge uniquement de la saisie des lexies.
- Le modérateur : en plus de la saisie des lexies, il se charge de la validation des lexies saisies par les membres de son groupe.
- Le super modérateur : se charge de valider les lexies saisies par l'ensemble des groupes.

L'architecture globale de la base de données se résume dans le schéma ci-dessous :



5.2. La page d'accueil

C'est la première page qui s'affiche lors de l'ouverture de la plate forme. Il suffit de mettre le nom de l'identifiant, le mot de passe et d'appuyer sur la touche valider pour accéder à la page suivante. Dans cette page, chaque membre a son identifiant et son mot de passe (*cf.* Figure 1).

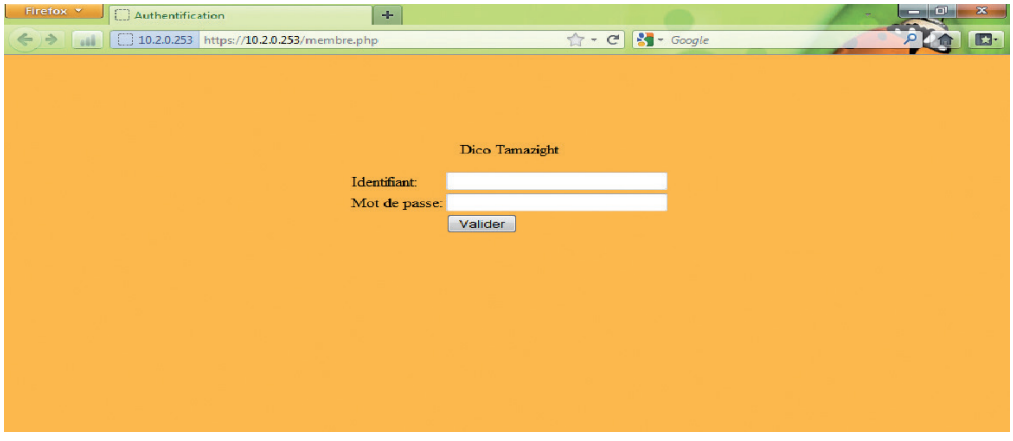


Figure 1 : Page d'accueil

5.3. Saisie des lexies

Cette rubrique, consiste à saisir la lexie ainsi que l'ensemble des informations qui la caractérise. En effet, elle nous donne des cases qui concerne, sa définition en tamazight, sa définition en français, sa transcription phonétique, sa racine, sa source (dictionnaire, usuel, manuel, terrain, etc.) et sa classe grammaticale (*cf.* Figure 2).

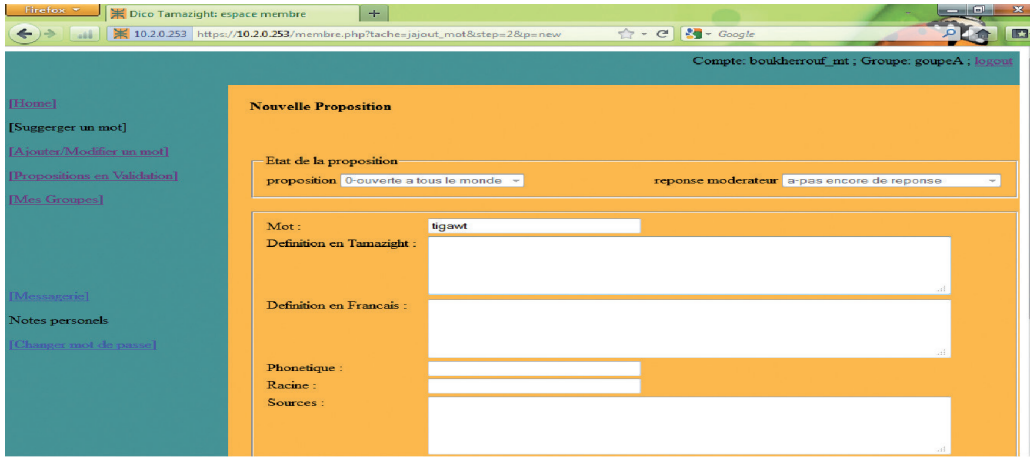


Figure 2 : Saisie des lexies

5.4. Présentation d'un exemple d'une lexie saisie

Dans ce point, nous présentons la forme de la page après la saisie des lexies. Nous donnons l'exemple de la lexie « tigawt ».

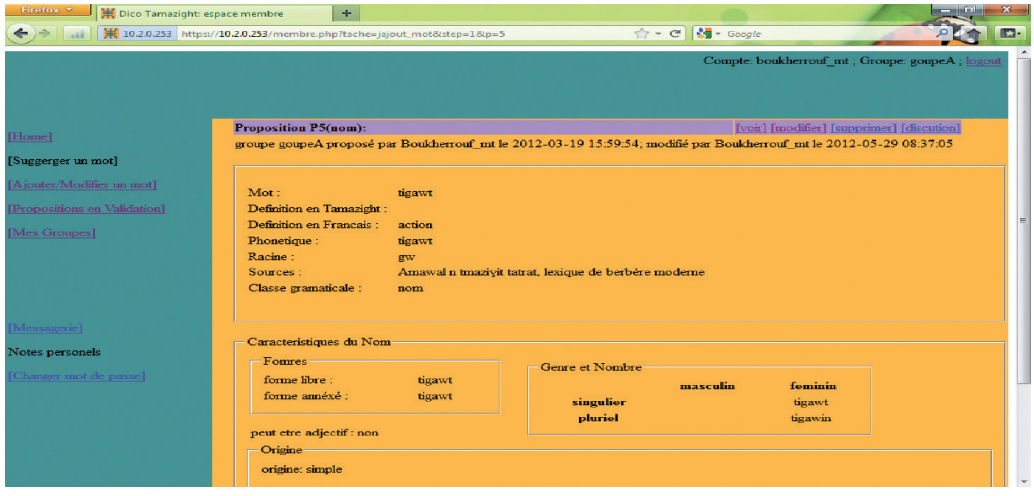


Figure 3 : Lexies saisies

Après avoir saisi la première page, la base nous oriente vers une autre page qui consiste à remplir les caractéristiques de la classe grammaticale en question. Etant donné que nous nous sommes limités à la terminologie de spécialité, l'écrasante majorité des lexies concerne la classe du nom. A cet effet, la base est alimentée par des informations concernant la catégorie du nom (genre, nombre, état, nom simple, nom dérivé, nom composé, etc.) (*cf.* Figure 3).

5.5. L'enregistrement des lexies

Dans la page réservée à l'enregistrement des lexies, nous trouvons toutes les informations relatives à la lexie. En effet, nous trouvons le numéro et la catégorie de la lexie, le groupe qui l'a saisi, le nom et le statut du membre qui a alimenté la lexie ainsi que la date, le membre qui l'a modifié ainsi que la date, en fin le statut de la lexie : valid. pour la lexie validée, ouverte pour la lexie qui n'est pas encore validée par le modérateur (*cf.* Figure 4).

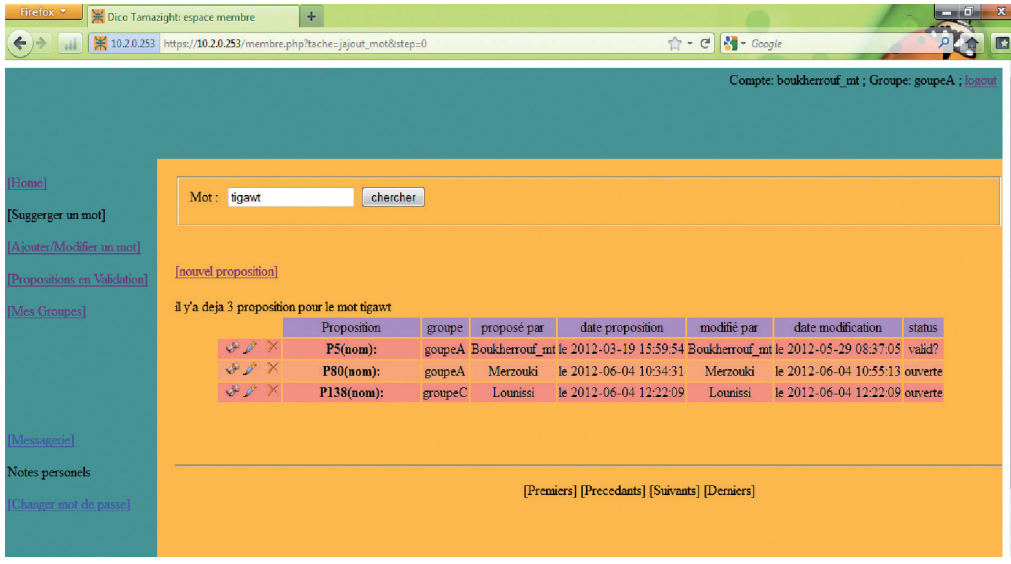


Figure 4 : Enregistrement des lexies

6. Conclusion

En guise de conclusion, nous pouvons dire que notre travail est limité à faire un inventaire et à regrouper toutes les propositions en matière de terminologie de specialité dipoinbles en tamazight dans toutes ses variétes. Notre base de données facilitera le travail de depouillement des aminageurs lors de leur intervention sur la tarminologie à adopeter.

Bibliographie

Amawal n tusnakt, *lexique de mathématiques*, Tizi-Ouzou, Tafsut, 1984.

Amawal (lexique), berbère-français), Paris, imedyazene, 1980.

Berkai A, (2002). *La terminologie linguistique en tamazight*, Magistère de berbère, Université de Béjaia.

Bouamara K., Rabhi A. (2000). *Amawal n tussna*, Université de Béjaia

Bouamara K. (2007). *Amawal n tunuγin n tesnukyest* « *Lexique de la rhétorique* », HCA, Alger.

Boumalk A., Nait Zerrad K. (2009). *Vocabulaire grammatical amazighe*, IRCAM, Rabat.

Chaker S. (1983). *Un parler berbère d'Algérie (kabyle), syntaxe*, Université de Provence.

Chaker S. (1991). *Manuel de linguistique berbère*, Ed. Bouchène, Alger.

Chaker S. (2003). *Atour de la racine en berbère : statut et forme*, Folia Orientola.

Chaker S.(1996). *Manuel de linguistique berbère –II : syntaxe et diachronie*, ENAG-Editions, Alger.

Cortade J-M.(1967). *Lexique français-Touareg, dialecte de l'ahaggar*, Paris, Arts et métiers graphiques.

Dallet J-M. (1985). *Dictionnaire français-kabyle : Parler des Ait menguellat*, Paris, SELAF.

Delheure J. (1987). *Dictionnaire Ouargli-français*, Paris, SELAF.

Foucauld Ch.(1951). *Dictionnaire touareg-français, Dialecte de l'Aheggar*, 4 tommes, imprimerie nationale de France.

Mammeri M. (1976). *Tajerrumt n tmazight*, Paris, François Maspero.

Manuels de L'éducation nationale (tamazight) primaire, moyen et secondaire.

Prasse K-G *et al*. (2003). *Dictionnaire Touareg- Français* (Niger), Press université de Copenhague.

Saad Bouzefran S. (1996). *Lexique d'informatique* (français, anglais, berbère), *amawal n tsenselkimt*, Paris, l'harmattan.

Salhi M-A. (2011). *Petit dictionnaire de littérature*, Odyssée, Tizi-Ouzou.

Taïfi M. (2002). *Dictionnaire Tamazight- français (parlers du Maroc Central)*, Paris, l'Harmattan.

Les Humanités Numériques et le Projet de Bibliothèque Numérique Franco-Berbère

Alain Vaucelle[1] Tassadit Yacine[2]
[1] Département ARTEMIS, TELECOM SudParis, Institut Mines-Télécom, Evry, France / Maison des Sciences de l'Homme Paris Nord, Saint-Denis, France
Alain.Vaucelle@telecom-sudparis.eu
[2] Laboratoire d'Anthropologie Sociale, Ecole des Hautes Etudes en Sciences Sociales, Paris France yacine@msh-paris.fr

Résumé

Le projet Bibliothèque Numérique Franco-Berbère (BNFB) est un projet financé par l'Organisation Internationale de la Francophonie. La BNFB a pour objectif de créer des dynamiques d'élaboration de ressources numériques franco-berbères normalisées.

Ce projet a pour ambition de fonctionner comme un portail documentaire de ressources berbères et de faire communiquer entre elles des bases de données diverses et hétérogènes, et donc de réaliser des partenariats entre plusieurs établissements que rapprochent leurs collections (complémentarité des fonds) ou leurs publics (services culturels d'une même collectivité).

Ce projet de corpus linguistique nous amène à nous interroger sur les méthodes des Humanités numériques. Cet article propose des éléments de réflexion sur les enjeux socio-économiques des processus numériques d'appropriation et d'organisation de la culture savante en Arts, Lettres et Sciences sociales par rapport aux nouveaux paradigmes du numérique.

1. Le projet BNBF

La BNFB a pour objectif de syndiquer et de mettre en réseau toutes les dynamiques d'élaboration de ressources numériques berbères. Ces ressources s'adresseront à la fois à des communautés académiques (revues savantes ou spécialisées, colloques, fonds libyques, fonds ethnolinguistiques et ethnomusicologiques), mais aussi à des communautés d'apprenants (formation initiale ou d'adultes, baccalauréat berbère), et au grand public.

En territoire touareg nigérien le projet se déploie sur un cycle d'ateliers pilotes d'enseignement traditionnel par les femmes de contes en français et en langue touareg. Les outils pédagogiques de ces ateliers seront à la fois classiques mais aussi multimédia (captation sonore et tablette graphique numérique).

La réalisation de contenus Franco-berbères (voire franco-arabo-berbères[1]) est stratégique pour la Francophonie car ces langues sont une composante fondamentale du monde francophone

1 Voire aussi de contenus multilingues prenant en compte des langues africaines sahéliennes

du Maghreb et du Sahel ainsi que de la diaspora vivant en Europe. Les langues berbères relèvent majoritairement de la culture orale (bien qu'il s'agisse d'une culture écrite plusieurs fois millénaire), ce qui rend difficile leur transmission intergénérationnelle avec des outils traditionnels alors que des solutions numériques multimédias innovantes seraient parfaitement pertinentes (capacité de numériser des corpus oraux, de les structurer, de les traiter, de les «feuilleter» à l'instar d'un texte). Nous proposons dans ce projet d'éditer numériquement des ressources en direction d'un lectorat et d'auditeurs (car nombre de ces ressources peuvent être sonores), disparates, et souvent pauvres. Ces ressources doivent pouvoir être déclinées sous forme orale ou écrite, sous plusieurs alphabets au choix : la coexistence de trois écritures (latine, arabe et tifinagh) correspondant à trois traditions justifie aussi ce choix, le passage de l'une à l'autre s'opérant potentiellement d'un clic de souris. Le choix du tout numérique est dicté par la dispersion des locuteurs (tant sur leurs territoires d'origine qu'en diaspora), par la pluralité des langues berbères se référant à un patrimoine commun : kabyle, rifain, chleuh, chaouia, touareg... mais aussi par l'émergence d'un marché en forte expansion (ordinateurs portables, tablettes, mais surtout ordiphones en anglais *smartphone* et livres électroniques).

La BNFB vise donc à rassembler les corpus numériques (textuels, audiovisuels, pédagogiques ou patrimoniaux) permettant de préfigurer les usages du futur mais aussi ceux des populations numériquement « branchées » qui grandissent d'années en années notamment chez les jeunes quelque soit leur origine sociale ou territoriale.

Les partenaires sur ce projet sont : Télécom SudParis (Département ARTEMIS-France), Alliance Cartago (France), AUF (Agence Universitaire de la Francophonie), Université Evry Val Essonne (Laboratoire Analyse & probabilité EA 2172, France), la Faculté des Sciences Humaines d'Agadir et l'Université d'Oujda (Maroc), l'Université de Tizi Ouzou (Laboratoire LAELA : Aménagement et Enseignement de la Langue Amazighe-Algérie), Université Abdou Moumouni de Niamey-Faculté des Lettres et Sciences Humaines-Département de Linguistique (Niger).

Ce projet s'adresse à la Communauté linguistique et la diaspora berbère pour des usages pédagogiques ou patrimoniaux, ainsi qu'à la communauté savante des études berbères et libyques et un public cultivé généraliste. Dans ce projet les jeunes sont ciblés en priorité car il est fondamental que l'enfant ayant le berbère pour langue maternelle puisse aussi utiliser des outils numériques adaptés pour déployer son intelligence. Bien que l'handicap d'appartenir à une langue dominée comme le berbère puisse déclencher parfois l'excellence intellectuelle, il s'avère que c'est souvent une cause d'échec. Des investissements relativement modestes dans la réalisation de ces fonds numériques, outils patrimoines et ressources d'enseignement en ligne sont à même de répondre aux besoins de ces jeunes. L'épreuve du baccalauréat berbère en France destinée à 2 à 3000 candidats (sous cinq langues berbères distinctes) est emblématique de ce besoin non satisfait pour des jeunes. Là aussi la syndication des ressources s'avère rentable pour répondre aux besoins des jeunes, mais aussi à ceux des adultes.

Notons que la cible adulte est particulière pour ce qui est des femmes souvent moins confrontées au monde extérieur, statistiquement moins alphabétisées[2] et pour lesquelles il est pertinent de pouvoir s'alphabétiser dans la langue qu'elles pratiquent, et au-delà, grâce au multilinguisme, dans les langues partenaires du berbère : le français mais aussi l'arabe et d'autres langues sahéliennes. Le cas du berbère en territoire sahélien est particulièrement ciblé dans le projet. Du fait de son statut matriarcal, la femme Touareg enseigne depuis des millénaires la langue et l'écriture (tifinagh). Alliant tradition et modernité ces ateliers de contes français et berbères sont expérimentés dans le Sahel nigérien.

2. Des enjeux sociétaux pour les langues berbères

La réalisation de grands corpus linguistiques est un enjeu sociétal, et cognitif fondamental. En effet l'accumulation de ces patrimoines numériques linguistiques conditionne la disponibilité de référentiels indispensables pour réaliser et optimiser la traductique du futur et le web sémantique et donc la présence et le référencement de ces corpus sur le web. Les populations au Maghreb, au Sahel ou en diaspora sont confrontées à un multilinguisme et un multiculturalisme beaucoup plus complexe que ceux des pays de large culture de l'imprimé. La situation culturelle linguistique du Maghreb est traversée par une richesse de disparités, des hiérarchies langagières complexes : langue maternelle, langue dominante du milieu, cultures orales, cultures écrites, polyglossies fonctionnelles : arabe dialectal/arabe classique, franco/arabe ou arabo-berbère, etc.

Cela pose donc des questions d'ordres stratégiques qui touchent à la compétitivité cognitive des individus et à la reconnaissance culturelle des populations. La réalisation pratique de grands corpus linguistiques numériques multilingues s'attachant à la fois à la collecte de corpus oraux et de leur transcription, leur traduction, et leur complexe interprétation. Le rassemblement numérique et interopérable de ressources manuscrites ou imprimées (voire pétrographiques) qui constituent le patrimoine commun du Maghreb sont des enjeux socio-économiques et culturels d'autant plus urgent que l'on a pu assister en Afrique à la destruction de patrimoines millénaires pour cause de conflits. Une des principales difficultés est de pouvoir numériser des ressources qui ne sont pas référencées au sein des fonds détenus par les grandes bibliothèques nationales (pas seulement au Maghreb) mais dans des centaines d'autres endroits, comme les bibliothèques du désert (Gaudio, 2002). La question dans ce projet est donc aussi de s'intéresser à la méthode de collecte associée à des méthodes des Humanités numériques que les projets pilotes à venir devraient idéalement intégrer.

Les méthodes des Humanités numériques ont pour objet de codifier les processus numériques d'appropriation et d'organisation de la culture savante en Arts, Lettres et Sciences sociales par rapport aux nouveaux paradigmes du numérique, de façon similaire à ce que leurs prédécesseurs les « Humanistes » avaient fait par rapport aux nouveaux paradigmes sociétaux et techniques qu'induisaient l'imprimerie. Ce sont donc non seulement les enjeux

2 A l'exception du contre exemple touareg, société matriarcale dans laquelle les femmes enseignent traditionnellement la langue et l'écriture tifinagh aux enfants.

sociotechniques des technologies de l'information et de la communication qui sont pris en compte, mais aussi les enjeux socio-informatiques dans l'appropriation des outils et des pratiques, la dissémination, et la pérennisation des résultats. Notre enjeu n'est pas, bien sûr, de redéfinir ce rôle des Humanités digitales, mais plutôt de définir un contexte. Notre ambition se limite à chercher concrètement et à affiner les conditions spécifiques d'appropriation des Humanités numériques dans un contexte multilingue et historiquement complexe que constitue le Maghreb et l'Afrique subsaharienne.

Les pays du Maghreb (et plus largement la communauté savante s'intéressant aux enjeux linguistiques précités) sont donc confrontés à un triple défi. Le premier est de pouvoir s'approprier les méthodes des Humanités numériques dans des communautés linguistiques et territoriales beaucoup plus complexes que la communauté académique anglophone. C'est cette communauté anglophone qui a inventé ce concept en s'appuyant sur un partage académique collégial qui s'organisait avec une culture largement unilingue et une histoire du territoire nord américain beaucoup plus consensuelle[3]. Le deuxième écueil réside dans l'appropriation de ces Humanités numériques en partageant certes la notion fondatrice d'Humanités, mais sans oublier que cette notion est nécessairement perçue comme un apport académique de l'excolonisateur, et non pas comme cela pourrait l'être idéalement comme un bien commun universel. Enfin, le Maghreb est de plus confronté à la réalité d'un historique culturel très ancien mais moins univoque que les Humanités européennes liées à la révolution guttembérienne de l'imprimerie, à la Renaissance européenne et à la culture latine.

3. Sciences de l'information et de la communication et Humanités numériques

Il est complexe d'analyser en termes de sciences de l'information et de la communication les nombreuses questions qui touchent à l'aménagement linguistique des grand corpus patrimoniaux numériques ; difficile aussi de trouver une légitimité des Humanités numériques et d'illégitimer parallèlement son refus par les acteurs scientifiques traditionnels des SHS et des Humanités. La raison en est la différence flagrante d'industrialisation des langues et d'appropriation académique des potentiels qui en dérivent : bibliothèques numériques, pratiques des Humanités numériques. De ce fait, les grands ensembles territorio-linguistiques sur lesquels se répartissent les chercheurs sont directement concernés et sont de ce fait, à la fois objets et sujets. Il est donc difficile pour des chercheurs anglophones ou ceux d'une grande langue européenne dans un pays développé, de comparer sereinement leurs états de l'art respectif et d'observer les niveaux d'appropriation pour ce qui est des bibliothèques numériques, des Humanités numériques et des enjeux de performance en ingénierie des langues qui y sont directement associés : interopérabilité normative des documents et des ressources, performance éditoriale, documentaire ou traductique.

La deuxième grande question vient de ce que, comme le décrit si bien Bertrand Gille (Gille, 1978), un nouvel ensemble technique ne s'installe vraiment qu'à deux conditions

3 Tout au moins en apparence au niveau académique car il est bien évident que le « melting pot » est bien sûr complexe sur le plan historique, linguistique et territorial.

consubstantielles : la mise en place effective d'un environnement technique cohérent (ici des bibliothèques numériques « intercompatibles » et normalisées[4]) et l'installation d'une technoculture correspondante (*i.e.* l'appropriation de méthodes des Humanités numériques[5]). C'est cette étape, émergence universelle de la mutation vers de nouveaux *habitus* méthodologiques et de nouveaux environnements techniques s'articulant avec les derniers progrès de l'information structurée et balisée, qui sont loin d'être complètement adoptés et installés dans le milieu académique des SHS et des Humanités et avec de surcroît des disparités importantes suivant les degrés de développement, les continents et les langues.

On constate ainsi que les bibliothèques virtuelles numériques ne s'imposent pas de facto en synergie avec les bibliothèques traditionnelles selon la même évidente synergie fonctionnelle qui s'est naturellement installée entre la téléphonie fixe et la téléphonie mobile. De même, les *habitus* académiques traditionnels dans les sciences humaines et la recherche littéraire (les Humanités), ne se transforment pas naturellement pour s'élargir et adopter en synergie les propositions des Humanités numériques. Des retards s'installent, confortés par des certitudes scientifiques ou professionnelles établies depuis longtemps, dans la plupart des disciplines de SHS (linguistique, recherche littéraire, ethnologie), souvent bloqués aussi par certains professionnels des bibliothèques ou même des informaticiens, voire même certains chercheurs appartenant plus aux mentalités des « anciens » contre celles des « modernes ».

Dans le monde anglophone, les grands progrès des Humanités numériques, des bibliothèques numériques, du Web sémantique, de la traductique du futur en émergence ont été précisément rendus possibles parce que des grandes Fondations Nord Américaines, des grandes bibliothèques du monde développé se sont coordonnées, ont investi des fonds considérables. Ces mêmes sponsors subventionnent encore aujourd'hui le cadre institutionnel et les équipes de chercheurs correspondant. Cela permet la mise en place de ce nouveau contexte technique. Cela offre aussi aux communautés de chercheurs en Humanités numériques un accompagnement depuis maintenant près de 20 ans pour s'approprier ces nouveaux paradigmes et réussir ainsi à constituer des corpus structurés (surtout anglophones) relativement conséquents.

Dans le monde francophone européen par exemple (celui dans lequel est écrit cet article) les Humanités numériques sont un concept émergeant encore bien flou. Il n'en est pas de même au Canada (même francophone). Et pourtant depuis 2007, la France s'est dotée du TGE Adonis[6] qui devrait lui permettre de mettre fin aux fausses polémiques interdisciplinaires et d'accélérer l'appropriation dans le milieu des SHS de méthodes intercompatibles et standardisées de traitement des documents et des corpus linguistiques.

4 Conformes notamment aux instructions de l'OCLC (Online Computer Library Center).

5 L'Alliance of Digital Humanities Organisations (ADHO) a adopté comme publication principale, le journal officiel de l'ALLC « Journal of Digital Scholarship in the Humanities » publié par les Oxford University Press et deux autres publications de portée mondiale « DHQ, Digital Humanities Quartely » et « Digital Studies / Le champ numérique ». Voir aussi une cartographie conceptuelle de l'univers des Humanités digitales disponibles sur : http://www.allc.org/publications/mind-map-digital-humanities

6 Le TGE Adonis est une plateforme de recherche en sciences humaines et sociales qui permet d'assurer l'accès et de préserver l'accès aux données numériques produites par les SHS. http://www.tge-adonis.fr

Car telle est bien la question centrale, depuis une trentaine d'années les scientifiques des sciences exactes et expérimentales ont mis en place une collégialité numérique mondiale. Les acteurs des SHS tardent à s'y mettre car ils acceptent beaucoup plus difficilement que des normes d'interopérabilité, de réusabilités et de diffusion de leurs documents (et surtout de leurs corpus de documents) puissent être mises en place. Il est cependant absolument clair que ces normes distinguent clairement 3 types de balisage ou d'aménagement en métadonnées: (a) référentiel et documentaire (b) structurel et formel (c) sémantique. Ces trois niveaux garantissent que le chercheur n'est aucunement soumis à une quelconque entrave de sa liberté de chercheur. Le champ sémantique lui est largement ouvert pour instrumentaliser numériquement ses hypothèses innovantes, par contre le champ référentiel ainsi qu'un minimum de structuration consensuelle avec les autres communautés de chercheurs de sa discipline lui permettront d'assurer la communicabilité, la réusabilité de ses documents et corpus donc la possibilité (comme en sciences expérimentales et exactes) de faire grandir au niveau planétaire de très grand corpus de documents (majoritairement des textes[7]), d'y accéder collégialement et donc de pouvoir développer des recherches en SHS au niveau mondial, notamment dans des disciplines comme la recherche littéraire. Notons que jusqu'à présent, ces chercheurs n'avaient pas d'autre moyen que de collaborer par l'échange de publications dans un contexte peu normalisé et donc très faiblement réutilisable. Certes, depuis les débuts de l'informatique des nombreux chercheurs en sciences humaines avaient eu l'idée d'informatiser des documents et des corpus. Ils avaient ainsi produit quantité de travaux passionnants : analyser des lexiques d'auteurs, structurer des corpus et par là conforter des hypothèses d'analyse structurale et/ou stylistique. Les linguistes avaient assez souvent produit des corpus très conséquents. Les archéologues, les historiens avaient produit d'importants travaux sur quantités de corpus plus originaux les uns que les autres. Cependant l'interopérabilité intrinsèque des documents associés aux couches successives d'analyse et de recherche n'étaient pas standardisée.

4. Langue et sémantique en devenir

La mise en œuvre de gros corpus de documents, notamment des textes de patrimoine littéraire, philosophiques, sociologiques, linguistiques, est souvent considérée comme très complexe et très couteuse pour les décideurs politiques et économiques. Pourtant il est certain que dans un futur proche l'efficience de la traductique et du e-sémantique dans telle ou telle langue sera très directement fonction du volume et de la diversité des grands corpus de ressources textuelles rassemblés et structurés de façon normative. Dès lors les enjeux deviennent beaucoup plus stratégiques car ces deux progrès prospectifs sont très importants pour assurer la prospérité économique, industrielle et économique des communautés linguistiques.

Nous en sommes actuellement, tant pour la traductique que pour l'e-sémantique à une véritable mutation des méthodes. Jusqu'à un passé très proche tant la traductique que l'e-sémantique fonctionnaient (et fonctionnent encore) grâce à des dictionnaires, des règles syntaxiques, des grammaires, dans des contextes relativement limités rassemblés par les équipes qui ont conçu

7 Mais pas seulement : les photographies, les documents filmiques ou audio peuvent être associés à ces corpus ainsi que des graphiques des formules mathématiques ou chimiques, des plans, etc.

ces outils techniques. Or, l'extrême efficacité du Cloud computing, les progrès fantastiques des ordinateurs tant en puissance de calcul qu'en rapidité, font qu'il devient de plus en plus réaliste de désambiguïser le sens des énoncés ou des traductions en le comparant au contexte exhaustif des occurrences dans la totalité d'un patrimoine linguistique. Mais pour proposer ces nouvelles opportunités, il importe que les ressources textuelles (mais aussi les ressources orales voire multimédias) présentes sur le Web, soient normalisés, interopérables et balisées d'un point de vue linguistique.

L'ISO TC37 (terminologie et ressources textuelles) développe ainsi des familles de normes pour qu'à termes la totalité de l'information textuelle (voire même des corpus audio) ne soit pas comme aujourd'hui des masses de documents structurées au minimum, n'ayant pour agencement que leur seule logique éditoriale propre, mais qu'elles puissent être structurées et balisées linguistiquement de façon intercompatibles et normalisées, et ce de façon automatique donc avec un minimum d'intervention humaine.

Mais pour que cette mutation « e-sémantique » et traductique puisse advenir, il est indispensable que les communautés linguistiques concernées (pour nous ici francophones, arabophones et berbérophone), prennent en compte le développement de ces progrès et réalisent concrètement de très grands corpus littéraires et linguistiques (tendant à l'exhaustivité des ressources), structurés (ou structurables) selon ces nouvelles normes de balisage linguistique.

5. La plateforme retenue pour le BNFB

C'est dans ce contexte des Humanités numériques que nous avons retenu la plateforme open source Omeka[8] comme solution de balisage (Dublin Core) et de diffusion de nos ressources. Cette plateforme est simple d'utilisation tant en gestion administrateur qu'en partage des contenus. Omeka est au croisement de logiciels spécialisés dans la gestion de collections comme Greenstone, et de la gestion de contenus comme Wordpress. Omeka est un projet développé par le Centre pour l'Histoire et les nouveaux Médias Roy Rosenzweig, et l'université George Mason[9].

Cette plateforme est dotée d'outils d'importation et de gestion de multiples formats (images, audio, vidéo, textes). Son interopérabilité est garantie par la l'importation et la gestion des métadonnées Dublin Core[10], et de l'OAI-PMH qui permet le moissonnage des ressources entrantes et sortantes, ainsi que la gestion de corpus TEI (Text Encoding Initiativ)[11].

La compatibilité avec Zotero[12], le logiciel de gestion des données bibliographique et des documents de recherche est assurée à l'aide d'un module d'extension. Plusieurs modules d'extensions sont à la disposition des utilisateurs afin d'optimiser la gestion et la diffusion des contenus.

8 www.omeka.org
9 http://www.gmu.edu/
10 http://dublincore.org/
11 www.tei-c.org
12 http://www.zotero.org/

Dans Omeka, le balisage Dublin Core peut se faire directement en ligne, lors de l'importation des données au sein de la plateforme. Le Dublin Core est un schéma de métadonnées générique. Ce schéma permet de décrire les ressources numériques à l'aide de 15 éléments de description formels (titre, créateur, éditeur), intellectuels (sujet, description, langue,) et relatifs à la propriété intellectuelle. Le Dublin Core est une norme internationale ISO 15836, reconnue par le W3C.

La plateforme est compatible avec l'Open Archives Initiative (OAI, initiative pour des archives ouvertes). L'OAI est un protocole facilitant l'échange et la valorisation d'archives numériques. A l'aide de ce protocole des fournisseurs de services peuvent moissonner les métadonnées sur les sites de fournisseurs de données. Il est ainsi possible d'utiliser un protocole OAI pour créer un outil de recherche simultanée moissonnant des données dans plusieurs bases de données bibliographiques et affichant le résultat dans une seule fenêtre de recherche (indépendamment du lieu physique des métadonnées). L'OAI a un moteur de recherche spécifique pour le contenu en libre accès, BASE. Archives Ouvertes *Protocol for Metadata Harvesting* - L'OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting) ou protocole OAI facilite donc l'échange de données entre des fournisseurs de données (par exemple des bibliothèques ou des musées...) et un fournisseur de service (qui peut être aussi une bibliothèque, un centre de documentation, un portail thématique ou local désirant rassembler des données). Ce protocole d'échange permet de créer, d'alimenter et de tenir à jour, par des procédures automatisées, des réservoirs d'enregistrements qui signalent, décrivent et rendent accessibles des documents, sans les dupliquer ni modifier leur localisation d'origine.

Grâce au protocole OAI, une bibliothèque agissant en tant que fournisseur de données a la possibilité d'offrir une visibilité accrue à ses documents, notamment à ses publications électroniques ou à ses fonds spécialisés. Réciproquement, en tant que fournisseur de service, une bibliothèque peut réaliser une base de données ou un portail documentaire dans son domaine de spécialité ou sur un thème quelconque, en collectant les données descriptives de ressources et documents de tous types, accessibles sur l'Internet dans des entrepôts OAI. Enfin, le protocole OAI permet de faire communiquer entre elles des bases de données diverses et hétérogènes, et donc de réaliser des partenariats entre plusieurs établissements que rapprochent leurs collections (complémentarité des fonds) ou leurs publics (services culturels d'une même collectivité).

6. Des protocoles et des normes pour des bibliothèques numériques

On voit bien que les questions d'Humanités digitales, d'e-sémantique et de traductique du futur sont étroitement liées au développement de très grandes bibliothèques virtuelles qui (mieux que le Web actuel), ouvriront « l'accès intelligent » à la quasi-exhaustivité des ressources textuelles et audio du monde entier. Mais on sait bien aussi que ce progrès ne se réalisera

pas aussi rapidement dans toutes les langues. Le risque est de voir certaines communautés linguistiques ne pas prendre en compte les enjeux de ces Humanités numériques. Le danger est grand de voir se créer de nouvelles fractures numériques.

Autant on pouvait considérer (bien à tort) que les Humanités numériques n étaient pas une activité scientifique prioritaire, autant la performance traductique et e-sémantique aura un impact beaucoup plus direct, économique, industriel, et sociétal sur le devenir du Maghreb multilingue.

La question actuelle est donc de savoir comment et quand le monde maghrébin s investira dans ces enjeux actuels. Les projets dans lesquels nous sommes engagés sont bien modestes, ne sont pas suffisamment larges pour dépasser le premier niveau de bibliothèque numérique et dessiner un cadre véritable d'Humanités numériques.

Ce démarrage maghrébin local nous apparaît indispensable car il permettrait à une collégialité Nord-Sud de se mettre en place : de nombreux documents appartenant d'évidence au Maghreb et aussi à la France existent par exemple à la BNF.

Ces nouvelles méthodes d'Humanités numériques, mais aussi pour la linguistique (et son aspect d'ingénierie pour l'e-sémantique et la traductique) suggèrent que non seulement elles dépendent d'une mobilisation de la société civile, académique et politique, mais aussi elles sont étroitement liées au cadre normatif d'intégration et d'interopérabilité de la numérisation des documents. L'enjeu est déterminant pour les pays en voie de développement afin de garantir un accès et un partage des ressources du savoir pour le plus grand nombre.

Bibliographie

Ben Henda M., Hudrisier H. (2012). Les normes et standards des TICE, des enjeux primordiaux pour le Sud. *Revue Frantice*, n° 4.

Ben Henda M. (2012). Vision historique, technique et prospective des systèmes d'information et de communication: interopérabilité normative globalisée. Mémoire de HDR sous la dir. de Roland Ducasse, Université Bordeaux III, à paraître.

Ben Henda M., Hudrisier H. (2009). Normalisation et terminologies multilingues pour les TICE. *In Forum Terminologique International*, Université de Sousse 20 au 23 novembre 2009.

Calvet L.-J. (1974). *Linguistique et colonialisme : petit traité de glottophagie*. Payot, Paris.

Fergusson C. A. (1959). Diglossia. In *Word*, 15 (3): 325-340.

Gaudio A. (sous la dir. de). (2002). *Les bibliothèques du désert* : Recherches et études sur un millénaire d'écrits. L'Harmattan, Paris.

Gille B. (1978). *Histoire des techniques*. Collection Encyclopédie de la Pléiade, Gallimard, Paris.

Hudrisier H. (2009). La nécessité d'adapter Internet à la mondialisation linguistique. In *Critique de la société de l'information* (coordonné par J.-P. Lafrance). Les Essentiels d'Hermès, CNRS éditions, Paris, pp. 115-134.

Hudrisier H., Ben Henda M. (2008). Cartago : une terminologie large langue de l'enseignement électronique à distance. In *Les outils d'aide à la traduction*. Séminaire de l'Union Latine, Bucarest, février 2008. Hudrisier H. (2011). Normalisation et prospérité multiculturelle. Chap. 3. In *La norme numérique ; Savoir en ligne et Internet*. Sous la dir. de Perriault, J. & Vagner, C. CNRS éd. Paris, pp. 63-87.

Hudrisier H., Ben Henda M. (2009). Enjeux normatifs des TICE de l'enseignement des langues dans le contexte arabo-berbère. *Colloque international Les TICE et les méthodes d'enseignement/apprentissage des langues*, CNPLET, Alger et Paragraphe, Paris, 30 mai-1er juin 2009, Tipaza, Algérie.

Hudrisier H., Ould Braham O., Saleh I. (2008). La numérisation (BNB) et le e-learning (workshop de Tipaza, Algérie, 28-29 mai 2008). In *Etudes et documents berbères*, N° 27, éd. La boite à documents, Paris, pp. 175-183.

Hudrisier H., Romary L. (2003). Le balisage normalisé des concepts et documents en liaison avec les normes de l'EAD. In *Colloque Normes & standards pour l'apprentissage en ligne*, Versailles, 19 mars 2003. En ligne (consulté le 16/03/04) : http://www.initiatives.refer.org/Initiatives-2003/_notes/_notes/henri.htm. Reprint in *Études et documents berbères*, N° 19, 20, éd. La boite à documents, MSH Paris Nord, 2004.

Hudrisier H., Vaucelle A. (2009). Technical and normative scenarios in the medium. In *International Preservation News*. n°47, May 2009, IFLA PAC, ISSN 0890-4960, http://www.ifla.org/files/pac/IPN_47_web.pdf

Organisation Internationale de Normalisation. *ISO 24610 : Structures de traits*, 2006.

Organisation Internationale de Normalisation. *ISO/FDIS 24611 : Cadre d'annotation morphosyntaxique, en développement*, en développement.

Organisation Internationale de Normalisation. *ISO 24613 : Cadre de balisage lexical*, 2008.

Organisation Internationale de Normalisation. *ISO 24614 : Segmentation des mots dans les textes écrits*, 2010.

Organisation Internationale de Normalisation. *ISO/FDIS 24616 : Plateforme d'informations multilingues*, en développement.

Marçais W. (1930). La diglossie arabe. In *L'Enseignement Public*, Revue pédagogique, tome CIV, n° 12, pp. 401-409, tome CV, pp. 20-39.

Ould Braham O., Hudrisier H. (2006). La bibliothèque numérique berbère. In *La langue française dans l'aventure informatique*. Colloques Lexipraxis 2005 & 2006, Paris, éd. AUF & AILF.

Ould Braham O., Hudrisier H. (2004). Le berbère et les nouvelles technologies de l'information. In *Études et documents berbères*, n° 19, 20, éd. La boite à documents, MSH Paris Nord, pp. 293-294.

Ould Braham O., Hudrisier H. (2008). Recueil et constitution de corpus oraux dans le domaine berbère, (Salle du CNPLET, Alger, lundi 2 juin 2008). In *Études et documents berbères*, n° 27, éd. La boite à documents, Paris, pp. 193-204.

Romary L., Hudrisier H. (2004). TEI : Text Encoding Initiative. In *Études et documents berbères*, n° 19, 20, éd. La boite à documents, MSH Paris Nord.

Suleiman S. M. (1982). *Jordanian Arabic between Diglossia and Bilingualism Linguistic Analysis*. Amsterdam.

Talmoudi F. (1984). *The Diglossia Situation in North Africa, a Study of Classical Arabic/ Dialectal Arabic Diglossia with a Sample Text in 'Mixed Arabic'*. Orientalia Gothburgensia, 8.

Vannini L., Le Crosnier H (dir.)(2012). Net.lang, réussir le cyberespace multilingue. éd. C&F, Paris.

Vaucelle A., Hudrisier H., Ben Henda M., Klett F. (2009). *ConvMPEG-SCORM : Rapport final & Livre Blanc, APO ISCC 2009 13/ 01/ 2009*. http://www.alain-vaucelle.fr/archives/945.

Vaucelle A., Hudrisier H. (2010). Langages structurés & lien social. In *Tic & société*, vol. 4, n°1, Interactivité. http://ticetsociete.revues.org/790.

TICAM'12, p. 39-53

Problématiques d'Usage et d'Intégration des Langues Peu Dotées dans le Web des Données Ouvertes (Linked Open Data ou LOD) Cas de l'Amazighe

Hammou Fadili
Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
hammou.fadili@cnam.fr

Résumé

Les langues peu dotées ont du mal à s'imposer dans le domaine des nouvelles technologies d'information et de communication. Ceci est dû aux retards et difficultés cumulés depuis plusieurs années empêchant leur intégration dans les nouveaux systèmes informatiques. Les problèmes rencontrés peuvent être d'au moins deux types : ceux liés au support des contraintes technologiques, puis ceux liés à la création et à la mise à disposition des contenus dans des formats compatibles. Dans le cadre du Web des données ouvertes (Linked Open Data ou LOD), de nouvelles difficultés s'ajoutent à celles qu'on rencontre dans les systèmes classiques. Le défi en plus, concerne les moyens à mettre en place pour créer, publier et relier à d'autres, les données afin de les rendre facilement accessibles et réutilisables par les humains et les machines. Et pour cela, on a besoin de solutions technologiques sophistiquées capables de formaliser, convertir et traiter automatiquement les contenus pour générer des données structurées, analysables et compatibles ; facilement intégrables dans le LOD. Dans ce qui suit, nous allons dans un premier temps, rappeler quelques notions importantes du LOD et les problèmes technologiques qu'on rencontre souvent dans le cas des langues peu dotées, dont « l'Amazighe ». Puis, nous allons passer en revue les difficultés du niveau de la gestion des contenus dans le même contexte, comme celles liées à la création, à la diffusion et aux usages. La partie suivante sera consacrée à la présentation de quelques recommandations concernant l'intégration des langues peu dotées dans le LOD avant de rappeler les perspectives et conclure le présent article.

1. Introduction

Les langues peu dotées ont du mal à s'imposer dans le domaine des nouvelles technologies d'information et de communication. Ceci est dû aux retards et difficultés cumulés depuis plusieurs années empêchant leur intégration dans les nouveaux systèmes. Les problèmes rencontrés peuvent êtes principalement d'au moins deux types : ceux liés au support des contraintes technologiques, puis ceux liés à la création et à la mise à disposition des contenus dans des formats compatibles. Pour assurer leur avenir dans le monde connecté de l'Internet et par conséquent, leur avenir tout court, les responsables et acteurs du domaine des langues peu dotées ont le devoir de s'activer pour remédier à ces problèmes afin d'assurer leur évolution

et leur modernisation. Parallèlement, ils doivent prendre en charge leur développement d'un point de vue scientifique pour pouvoir y véhiculer la production, la traduction et la diffusion des savoirs entant que langues savantes, aspect faisant défaut et constituant un grand handicap aujourd'hui.

Dans le cadre du Web des données ouvertes (Linked Open Data ou LOD), de nouvelles difficultés s'ajoutent à celles qu'on rencontre dans les systèmes classiques. Le défi en plus, concerne les moyens à mettre en place pour créer, publier et relier à d'autres, les données afin de les rendre facilement accessibles et réutilisables par les humains et les machines. Et pour cela, on a besoin de solutions technologiques sophistiquées capables de formaliser, convertir et traiter automatiquement les contenus pour générer des données structurées, analysables et compatibles ; facilement intégrables dans le LOD.

Dans notre cas, nous avons souhaité mettre l'accent sur certains aspects qu'on peut généraliser à d'autres langues peu dotées par rapport aux exigences du LOD. Ce sont des difficultés rencontrées suite à une expérience de mise en place d'une brique essentielle de la préparation et de l'intégration d'une langue peu dotée dans le Web des données ouvertes, à travers, entre autres, une tentative de mise en place d'une ontologie scientifique en langue Amazighe.

Dans ce qui suit, nous allons dans un premier temps rappeler quelques notions importantes du LOD et les problèmes technologiques qu'on rencontre souvent dans le cas des langues peu dotées, dont « l'Amazighe ». Puis, nous allons passer en revue les difficultés du niveau de la gestion des contenus dans le même contexte, comme celles liées à la création, à la diffusion et aux usages. La partie suivante sera consacrée à la présentation de quelques recommandations concernant l'intégration des langues peu dotées dans le LOD avant de rappeler les perspectives et conclure le présent article.

2. Aspects technologiques du Web des données ouvertes (Linked Open Data ou LOD)

Le Web des données ouvertes est une notion récente, qui se popularise au niveau de la publication et de la diffusion des données « structurées » sur le Web. Il est considéré par certains comme une implémentation du Web sémantique et par d'autres comme une évolution pragmatique du Web sémantique (très vaste et très difficile à maitriser). Dans ce qui suit, nous allons rappeler brièvement la notion du Web sémantique suivie de la définition du LOD avant de décrire les problèmes que rencontrent, souvent, les langues peu dotées dans ce nouveau mode de publication Web.

2.1. Le Web Sémantique

Le Web sémantique est une notion très importante de l'Internet moderne, permettant non seulement le stockage et la diffusion de données, mais également leur analyse et leur compréhension par des raisonnements sur leurs sens, par des machines ou agents logiciels. Le Web sémantique est basé sur des standards comme RDF et OWL pour structurer et annoter les contenus sur lequel on définit et on construit des technologies permettant aux machines d'effectuer des traitements automatiques (dans certains cas difficiles pour l'homme, par exemple : indexation, compréhension automatique et recherche sémantique) en s'appuyant sur des concepts comme, l'expression du sens, La représentation et la gestion

des connaissances, Les ontologies, les agents, etc. (Tim Berners-Lee). Ci-après, un bref rappel des couches constituantes du Web sémantique reposant sur ce qu'on appelle la pile du WEB sémantique composée d'une hiérarchie de langages normalisés.

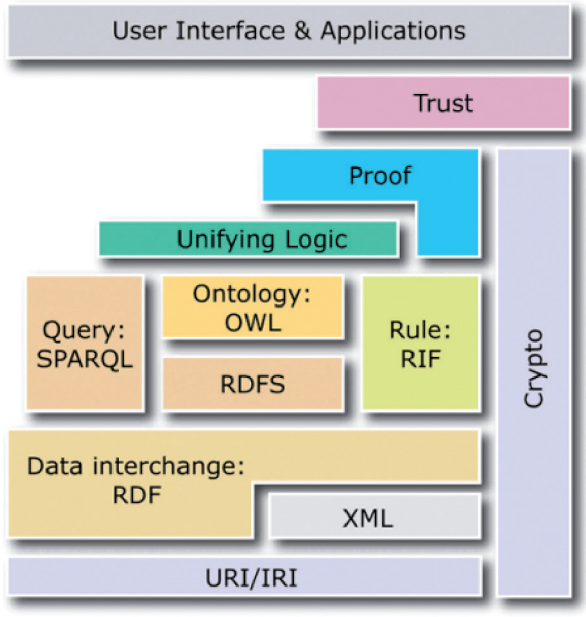


Figure 1: Pyramide du Web sémantique, Berners-Lee

Les langages de cette pyramide peuvent être regroupés pour former des couches cohérentes correspondantes à des niveaux de traitements et d'utilisations :

- Localisation des données, URI/IRI,
- Format de description des données et de résolution des espaces de noms, XML/S,
- Données ou description des données, RDF,
- Définition et description des schémas et vocabulaires des ontologies : RDFS, OWL, RIF, SPARQ,
- Logique, déduction de nouvelles règles, non prévues au départ, SWRL,
- Preuve, définition des outils de description des étapes du raisonnement logique,
- Authentification, définition des outils et des services d'authentification des données,
- Utilisateur, définition des interfaces et applications utilisateur.

2.2. Le Web des données ouvertes (Linked Open Data ou LOD)

Une donnée est un élément fondamental décrivant un fait brut non interprétée (devient une information une fois interprétée). Une donnée ouverte est une donnée mise à disposition dans des formats ouverts et normalisés permettant son exploitation sous des licences ouvertes, libres et gratuites garantissant sa réutilisation dans un contexte international.

2.2.1. Définition

> *« Le Web des données (Linked Data, en anglais) est une initiative du W3C (Consortium World Wide Web) visant à favoriser la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations. »* Wikipédia.

Le concept de données liées (Linked Data) a pour but l'exploitation du Web par les machines et par les êtres humains en faisant abstraction aux spécificités technologiques de toutes les sources de données liées. Celui des données ouvertes (Open Data) consiste à répondre à la nécessité de disposer de données légalement "ouvertes", c'est à dire librement (ré-) utilisable par l'utilisateur et pour n'importe quel but. Le Linked Open Data (LOD) est la fusion des deux concepts. Son but est de permettre l'exploitation du Web par des machines et par les êtres humains d'une manière libre, faisant abstraction aux spécificités technologiques de toutes les sources de données. C'est un moyen puissant pour connecter les utilisateurs à la connaissance et les utilisateurs entre eux **chacun dans sa langue** et dans des environnements conviviaux.

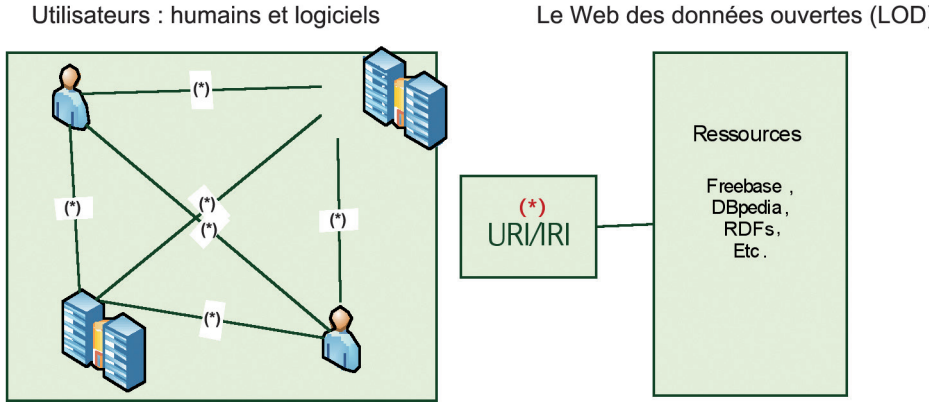


Figure 2 : Les utilisateurs du LOD

C'est une évolution du Web actuel (Web 2.0) où les utilisateurs interagissent avec des sources de données fixes, vers un « Web de données », où les utilisateurs et les machines interagissent potentiellement avec le Web entier pouvant être considéré comme une seule base de données universelle, composée des sources de données liées et compatibles. C'est une notion récente et très importante de l'Internet moderne qui se popularise, essentiellement au niveau des grandes institutions et des gouvernements, pour la publication de leurs données publiques

dans le cadre de la bonne gouvernance, transparence, etc. Dans le cadre de cette évolution du Web, Tim Berners-Lee[1] a défini quatre principes sur lesquels doit reposer l'intégration des données dans le LOD :

1. On doit utiliser les URIs comme noms des ressources.
2. On doit utiliser les URIs HTTP accessibles aux utilisateurs.
3. On doit utiliser les normes et standards RDF et SPARQL pour la description et l'interrogation des données.
4. Et puis, on doit créer des liens vers d'autres URIs afin de permettre de découvrir de nouvelles ressources.

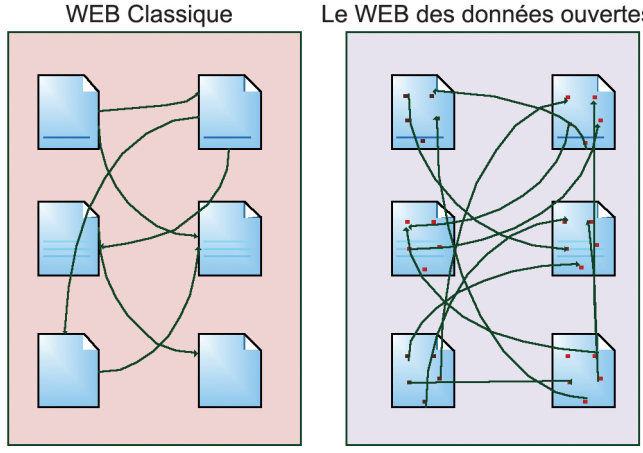


Figure 3 : Evolution du Web vers le LOD

Contrairement au Web classique permettant, seulement la création de liens entre les pages, le Web des données ouvertes permet, en plus, la création de liens entre les données à l'intérieur des pages. Et pour mesurer la qualité de la publication de ces données, Tim Berners-Lee a donné également cette classification en nombre d'(*).

1 http://www.w3.org/DesignIssues/LinkedData.html

Les conditions associées à chaque ligne sont, à cumuler en plus, de celles des lignes qui précédent.	
★	Publication des données sous n'importe quel format, mais avec une licence libre (données ouvertes).
★ ★	Publication des données structurées (RDF, Excel, XML, CSV, etc.).
★ ★ ★	Utilisation des formats non propriétaires (exclure Excel et tous les formats propriétaires).
★ ★ ★ ★	Utilisation des standards du W3C pour identifier les ressources (URI, RDF, SPARQ, etc.).
★ ★ ★ ★ ★	Création de lien vers d'autres ressources (définition du contexte des données).

Figure 4 : Qualité des données dans le LOD

2.2.2. Principaux langages du LOD

L'interopérabilité est un des éléments principaux du Linked Open Data. En effet, les informations isolées n'ont aucune valeur : une donnée a de l'importance quand elle est liée à d'autres. Car, elle peut être atteinte à partir d'autres, être exploitée, réutilisée, etc. Pour assurer l'interopérabilité avec l'existant, dans le but de faciliter la conversion et la publication des données existantes dans le LOD, la plupart des formats, techniques, méthodes et outils (RDF, SPARQL, OWL et RDFS) actuels y sont supportés. C'est plutôt une bonne chose, du fait que cela repose sur des normes déjà utilisées, qui vont aider à intégrer et à réutiliser l'existant à moindre coût. D'ailleurs, cette vision sur la manière de faire évoluer le Web actuel vers le LOD est soutenue par les grandes institutions et fournisseurs de contenus, car cela va leur permettre de publier facilement leur données à partir de leur format initial. En effet, rendre les données non conformes, compatibles et structurées suivant un standard donné nécessite des traitements qui peuvent dans certains cas coûter plus cher.

Les langages constituant le LOD permettent la gestion de la base de connaissances et les raisonnements associés pour permettre d'une part, une meilleure présentation des connaissances pour les utilisateurs humains et d'autre part, une meilleure représentation des connaissances pour les machines. Ci-après, des éléments de description des principaux langages, sous-ensemble de la pyramide du Web sémantique.

- Identification & nommage

URI : les URIs (Uniform Resource Identifier) permettent d'identifier d'une manière unique les ressources sur le WEB.

IRI : les IRIs (Internationalized Resource Identifiers) permettent aux personnes d'identifier des ressources Web dans leur propre langue.

- Description des données

XML : Extended Markup Language (XML) est un langage de description et d'échanges de documents et des données ne permettant pas leur présentation, appelé aussi format d'échanges standardisé. Il permet aux applications reconnaissant ce format d'échanger tous les types de données décrits dans ce langage, utilisé souvent pour assurer la compatibilité des données entre les applications hétérogènes.

RDF : Resource Description Framework (RDF) est un métalangage qui sert à décrire les ressources, leurs propriétés et les valeurs des propriétés sous forme d'un graphe (ressource, propriété, valeur). Il est considéré comme un modèle standardisé de description des métadonnées qu'on peut définir et associer à des documents ou de description des annotations qu'on peut définir et associer à des éléments d'un contenu. Ces annotations et métadonnées permettent d'associer du sens à des contenus qui peuvent être traités d'une manière automatique par les agents logiciels. Pour la gestion sémantique des données, les métadonnées RDF peuvent être des informations sémantiques associées à des mots du texte. Ces éléments peuvent être ensuite analysés et interprétés pour l'extraction du sens global. A noter que RDF peut être exprimé dans plusieurs langages, mais c'est XML qui est souvent utilisé, on a crée pour cela une version XML de RDF appelée RDF/XML.

- Sémantique des données

RDFS : RDF peut être aussi utilisé pour décrire des situations et des utilisations particulières avec un vocabulaire bien précis en utilisant la notion de RDF Schéma (RDFS). RDFS consiste à adapter RDF à des domaines modélisés particuliers décrivant des utilisations particulières au sein d'une communauté. On peut associer une ontologie de métadonnées ou d'annotations partagées définissant le contexte d'utilisation et échangeables entre les différents agents humains et logiciels. La définition d'un schéma RDF consiste en une activité de typage et de classification des ressources, des propriétés et des relations sous forme de classes définies dans des « espaces de noms » servant principalement à désambigüiser les mêmes éléments d'un vocabulaire définis dans des utilisations ou espaces de noms différents.

OWL : OWL est une extension de RDF enrichie avec des propriétés sémantiques, de contraintes, de comparaisons, de cardinalités, etc. pour décrire et manipuler les ontologies. C'est un langage recommandé par le W3C basé comme RDF sur XML qui permet à des moteurs d'inférences d'agents l'interprétation et le raisonnement automatique sur les ontologies. En effet, OWL est basé sur les logiques de description utilisées dans les systèmes de représentation de connaissances et offrant de fortes possibilités de manipulation de prédicats de classes, de rôles et d'individus, donc d'ontologies, contrairement aux logiques de premier ordre classiques ne manipulant que des objets de même type.

- Interrogation

SPARQL : SPARQL (Simple Protocol and RDF Query Language) : Langage d'interrogation des ontologies représentées sous forme de graphes RDF/S. Il est pour les bases de connaissances RDF ce que SQL est pour les bases de données relationnelles. Exemple :

SELECT ?x, ?y, ?z FROM URI/IRI WHERE {Conditions1, condition2, etc.}

2.2.3. L'évolution du LOD

Le LOD a un succès grandissant ces dernières années, lui permettant de s'enrichir continuellement. On peut constater son évolution sur une année, entre 2009 et 2010, sur les images suivantes (extraites du Cloud du LOD) montrant l'ensemble des données publiées suivant les formats et principes du « Linked Open Data ». Des initiatives pour rejoindre le LOD voient le jour régulièrement ; parmi les principaux projets crées autour de cette notion, il y avait dans un premier temps les projets fondamentaux comme : le «Linked Open Data» du W3C, DBpedia sur la structuration et l'annotation des données semi-structurées de Wikipedia ou encore feebase sur l'agrégation et l'annotation collaboratives des données à partir de sources de données hétérogènes ; puis dans un deuxième temps, les projets des institutions gouvernementales initiés principalement dans le cadre des démarches de la bonne gouvernance, la transparence, etc. ; suivies après par les grandes entreprises.

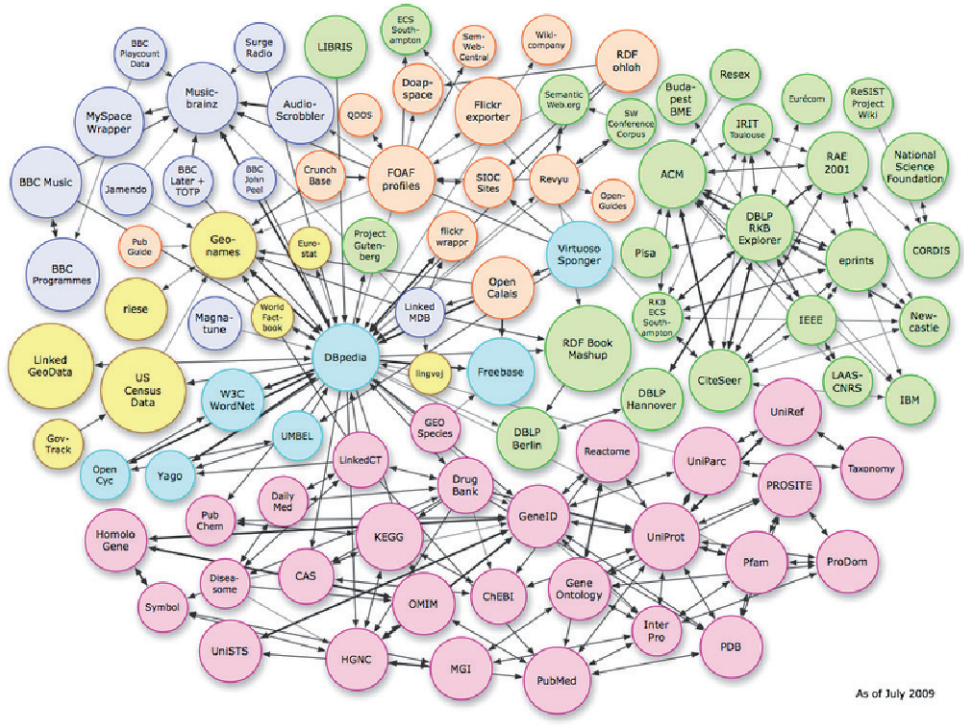


Figure 5 : Le Cloud du LOD en 2009

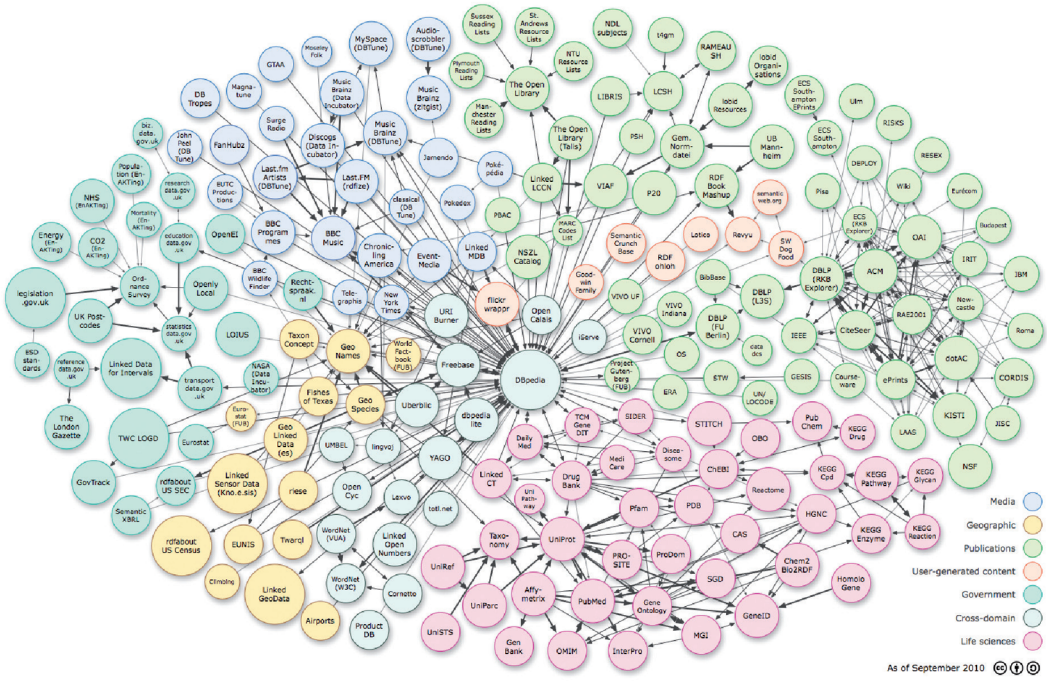


Figure 6 : Le Cloud du LOD en 2010

Le LOD est composé d'ensembles de données structurées et liées formant à leur tour un ensemble de données structurées potentiellement liable à d'autres ensembles et ainsi de suite (d'une manière récursive). En se basant sur ces éléments, plusieurs études ont été faites faisant le parallèle entre les données liées et les bases données relationnelles, montrant que l'évolution du Web des données (structuré, annoté, etc.) peut se comporter à terme comme une seule base de données universelle.

3. Exigences technologiques relatives aux traitements d'intégration des données dans le LOD

Dans cette partie, l'objectif est de relever les exigences et les problèmes à résoudre par les informaticiens et spécialistes de la langue permettant l'intégration des données dans le LOD. Ces problèmes on les rencontre d'une manière accrue dans le cas des langues peu dotées dont Tamazight sur laquelle des tests ont été effectués.

Dans le Web de données ouvertes, tous les types de données (pages, sections, éléments de bases de donné, données quelconques, etc.) doivent être basés sur des éléments de description et des liens sémantiques utilisables durant tout leur cycle de vie et leur présence sur Internet. Leur intégration nécessite la constitution de corpus normalisés structurés et annotés. Le cas des données structurées pose moins de problèmes que celui des données non structurées nécessitant des technologies sophistiquées, de bonnes pratiques et d'outils évolués pour effectuer les différentes tâches de traitements permettant

de les convertir et les rendre compatibles. D'une manière générale, on peut distinguer au moins trois catégories de données : les données structurées de type bases de données, RDF, etc., les données semi-structurées (qui ont une structure irrégulière par balisage) de type XML et compatibles et enfin les données non structurées de type PDF, texte, etc.

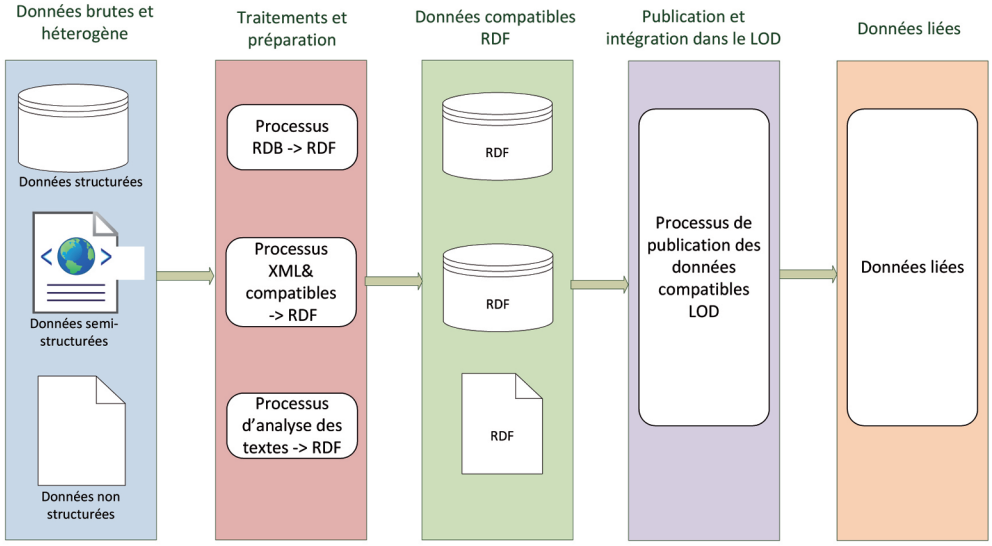


Figure 7 : Processus d'intégration des données dans le LOD

Une des plus importantes recommandations du LOD, concerne la prise en charge de tous les types de données, afin que l'intégration de l'existant soit la plus large possible. Ceci a bien évidement un impact sur le choix des technologies et des solutions à adopter. D'un point de vue traitements et pour simplifier le nombre de processus, nous pouvons distinguer et retenir, seulement deux types de données à qui on peut associer deux types de traitements : Les données structurées d'un côté, comme par exemple, dans le cas des bases données, les données de type simple (chaine de caractères, dates, nombres, etc.) et les traitements orientés « schéma » de leur structure ; puis de l'autre côté, les données non structurées ou les données complexes (textes par exemple) avec les traitements orientés « données » basés sur l'analyse sémantique, l'annotation, l'extraction d'informations pertinente à partir des textes, etc. Les données semi-structurées quant à elles peuvent être considérées comme une combinaison des deux.

3.1. Données structurées

Dans le cas des données structurées, le processus d'intégration est caractérisé par un ensemble de spécifications de critères et de règles portant sur le schéma et permettant de : récupérer les données structurées que l'on souhaite publier dans le Web des données ouvertes. D'une certaine manière leur intégration est facile à mettre en Suvre, car ne nécessitant pas

de traitements particuliers supplémentaires coûteux. Dans le cas des langues peu dotées, le problème concerne la constitution des données elles même, qui sont souvent très peu présentes ou inexistantes.

3.2. Données non structurées

Dans le cas des données non structurées, qui constituent, d'ailleurs, la majorité des données existantes et produites sur Internet (d'où la nécessité de leur intégration pour enrichir le Web des données ouvertes), un des problèmes importants de leur intégration concerne leur conversion dans des formats appropriés. Pour cela, on a besoin de s appuyer sur des outils et des éléments linguistiques autour du Traitement Automatique du Langage Naturel ou TALN pour l'analyse et la formalisation de la sémantique des documents permettant des traitements automatiques par des machines, comme le préconise le LOD. La linguistique est donc d'une grande importance pour le Web des données qu il faut continuer à développer au service de cette notion du futur. On peut déjà remarquer que beaucoup de travaux sont en cours sur ce sujet[2, 3] et des groupes de travail se sont également constitués comme l'Open Linguistics Working Group (OWLG)[4] qui a pour but de permettre de recueillir des études de cas, des recommandations et de bonnes pratiques à l'égard de la formalisation de la sémantique pour le LOD entre autres. C'est un domaine où les langues peu dotées ont beaucoup de retard, ce qui constitue un des obstacles supplémentaires pour leur intégration.

4. Exigences liées à la constitution de corpus

Dans cette partie, l'objectif est de relever d'une manière générale les problèmes à résoudre par les spécialistes fournisseurs de contenus et autres pour la réalisation de corpus informatisés et normalisés. Ces problèmes on les rencontre d'une manière accrue dans le cas des langues peu dotées dont Tamazight.

La constitution d'un corpus consiste à recueillir des informations de différents types à partir de différentes sources par rapport à un domaine donné. La création et la diffusion d'un contenu, doit respecter un certain nombre de critères et de méthodologies bien étudiées. Il faut se baser sur des études de cas, des recommandations, de bonnes pratiques, etc. fournies par les fournisseurs de contenus, d'outils et de solutions, assurant une veille dans le domaine. Il faut également, faire le bon choix des formalismes de représentation, des formats et des standards à adopter en rapport avec les types de traitements sur les données à stocker, à visualiser et à requêter.

2 http://linkeddata.org/
3 http://ldl2012.lod2.eu/
4 http://linguistics.okfn.org/201120/05//the-open-linguistics-working-group/

Dans le cas des langues peu dotées, ces critères posent des problèmes, en plus d'autres liés à leurs spécificités, comme par exemple pour :

- Le recueil des données à partir de sources sonores : les sources informatisées sont rares et quand elles existent, elles sont sous forme sonores, car ce sont des langues qui sont souvent de tradition orale. Se pose donc le problème de leur transcription et leur traitement afin qu elles soient exploitables.
- La définition des termes scientifiques : un autre problème important consiste en la création de termes scientifiques ou savants par rapport à un domaine ou à une discipline donnés : beaucoup de retard a été cumulé par rapport aux langues savantes. Dans le cas de la mise en place d'une ontologie scientifique, en amazighe, beaucoup de problèmes ont été rencontrés pour trouver et définir les mots « scientifiques » par rapport au domaine étudié.
- Etc.

D'autres problèmes liés aux langues peu dotées seront exposés dans le paragraphe «Recommandations pour l'intégration des langues peu dotées dans le LOD : le cas de l'amazighe».

5. Exigences liées à la problématique des usages et des copyrights

Dans cette partie, l'objectif est de relever d'une manière générale les questions à prendre en compte par tous les acteurs les institutions et autres pour la constitution, la diffusion et l'exploitation des données dans le LOD d'une manière légal et sans restriction. Comme pour les précédentes exigences, ces questions on les rencontre d'une manière accrue dans le cas des langues peu dotées dont Tamazight.

Le LOD est basé sur des données totalement formalisées qui ont beaucoup d'importance par le biais des liens et des partages. C'est la raison de l'adoption rapide de ce nouveau concept du Web par les institutions et les gouvernements qui souhaitent améliorer la transparence. Or, comme vu précédemment, l'intégration qualitative des données doit respecter les critères de Tim Berners-Lee dont celui de la gratuité. C'est-à-dire, pour que les données soient intégrées et proprement utilisées, il faut qu elles soient publiées sous « Open Licence ». C'est une exigence importante adressée aux créateurs et aux fournisseurs de contenus. Ce problème se pose moins dans le cas des grandes institutions et des gouvernements qui disposent déjà de grandes quantités de ce genre de contenus et qui souhaitent le publier dans le cadre de la bonne gouvernance. Cependant, il se pose d'une manière plus accentué dans le cas des langues peu dotées qui ne sont pas, parfois, « effectivement » officielles dans leurs pays.

Une autre exigence, concerne, l'utilisation et l'exploitation des ressources dans le contexte international. Dans le cas des langues peu dotées, ceci pose plusieurs problèmes du fait qu on ne dispose pas d'outils et d'applications capables d'aider et d'acheminer l'utilisateur vers la bonne information à partir d'autres langues : problème de traduction automatique lié aux inégalités des équipements informatiques dont souffrent les dites langues. Pour remédier à ce problème, il faut mettre en place des solutions basées sur des dictionnaires, correcteurs, traducteurs, etc. et de suffisamment de ressources normalisées (corpus, dictionnaires, ontologie, etc.) permettant d'assurer leur interopérabilité avec d'autres langues dans le nouveau Web mondial et multilingue.

6. Recommandations pour l'intégration des langues peu dotées dans le LOD : le cas de l'amazighe

Une partie de ces éléments est issue d'une expérience menée pour créer un corpus en amazighe intégrable dans le LOD. Pour cela, nous avons essayé de construire une ontologie scientifique en utilisant un système de gestion de bases de connaissances ontologiques (élément important du Web des données) dans le Cloud. Notre souhait était de doter Tamazight d'une brique supplémentaire : de ressources et contenus structurés et normalisés suivant les nouveaux critères du Web. Au cours de la réalisation de ce projet, plusieurs difficultés liées à la langue ont été rencontrées ; ceci a bien évidement un impact négatif sur le projet qui a retardé sa finalisation. Dans ce qui suit un récapitulatif des éléments faisant défaut pour l'évolution et l'intégration de cette langue dans les nouveaux systèmes d'information en général et dans le Web des données ouvertes en particulier. Ce sont bien sûr des éléments handicapants pour son développement.

Formalisation : la formalisation d'une langue consiste à définir et à mettre en place tous les modèles linguistiques permettant tous les traitements automatiques. Ceci se fait en général par la mise en place de règles formelles permettant l'analyse morphologique, syntaxiques, sémantiques, etc. définissant les modèles de représentation de la langue. C'est une étape importante pour le développent et l'intégration de toute langue dans les systèmes informatisés. Dans le cas de Tamazight, et jusqu à présent, plusieurs travaux ont été réalisés et d'autres sont en cours, mais le chemin reste encore plus long pour disposer d'une modélisation satisfaisante et complète de la langue. Cette première brique dans le traitement automatique de l'amazighe est très importante ; accentuer les efforts dans ce domaine devient une nécessité absolue.

Outils du TALN : un autre point en rapport avec le premier, concerne le développement et la mise en place d'outils de traitement automatique de l'amazighe (TAL-A) basés sur les modèles de la langue pour développer des systèmes informatiques capables d'analyser, de reconnaître, d'interpréter et de reproduire le langage naturel sous ses différentes formes. C'est un domaine difficile qui demande beaucoup d'efforts. Et la constitution d'une branche solide autour de ce sujet est nécessaire ; réunissant et rapprochant les spécialistes du domaine y compris ceux d'autres langues, afin de profiter des avancements faits ailleurs et les adapter pour la langue amazighe.

Contenus : Tamazight est une langue de tradition orale, les contenus écrits y sont rares, excepté ces dernières années ou la production a un peu évolué, mais reste très faible. Pour sa modernisation, Tamazight a un double défi : Constitution de contenus, puis leur formalisation et structuration. D'une certaine manière si Imazighen souhaitent sauvegarder leur langue, il faut absolument produire et produire. Pour cela, toutes les initiatives pouvant aider à encourager la création de nouveaux contenus sont les bienvenues, comme par exemple : le traitement des contenus de types autres que texte (audio, vidéos, etc.) constituant une grande partie des documents aujourd'hui ; et pour cela, on a besoin de nouveaux outils de reconnaissances de la parole et de transcription pouvant aider à la génération de nouveaux contenus. Ou, la participation à des initiatives internationales pour le développement de solutions et la constitution de ressources partagées pour le Web de données ouvertes comme le Open Linguistics Working Group (OWLG).

Normalisation : comme décrit précédemment, une donnée isolée est une donnée peu ou pas intéressante. Ceci est aussi vrai pour les langues : une langue isolée est une langue peu ou pas du tout intéressante notamment au niveau du Web. Afin d'assurer une présence effective de l'amazighe sur Internet : interopérable, interrogeable, exploitable, partageable, etc., il faut suvrer pour sa normalisation et la normalisation de ses outils et de ses contenus. Ceci afin d'assurer son intégration et sa diffusion dans le LOD : agrégation de contenus écrits dans plusieurs systèmes hétérogènes, traductions vers/à partir d'autres langues, etc. Pour cela, il faut tout d'abord créer des contenus structurés et annotés basés sur des normes internationales. Comme la mise en place d'ontologies (vocabulaire commun partagé par les spécialistes d'un domaine) qui ont le double avantage : d'une part, elles constituent la base du Web des données ouvertes et d'autre part, elles permettent de développer le côté savant de la langue. En effet, la constitution d'une ontologie est une concertation entre les spécialistes pour modéliser un domaine scientifique donné. Dans le cadre de notre expérience sur la constitution d'une ontologie scientifique, trouver les termes associés aux concepts et aux relations scientifiques est un chalenge absolu. Ceci pourrait bien sûr être, un élément bloquant pour la constitution de corpus scientifiques dans les langues peu dotées en général. Pour l'intégration des données non structurées, il faut mettre en place des systèmes d'analyse des textes, évolués et intelligents permettant la création et la génération d'annotation, nécessaires pour l'analyse automatique de contenu, l'extraction/fouille sémantique des données, traduction, etc. C'est le seul moyen d'intégration des données non structurées dans le Web de données ouvertes.

Graphie : afin de profiter de la réutilisation des applications et plateformes existantes dans d'autres langues, qui pourrait constituer un gain (temps, moyen, etc.) pour le développement de l'amazighe, il faut Suvrer pour le support de Tifinagh. Il faut prendre des initiatives auprès des éditeurs de solutions et d'applications pour les inciter à prendre en considération le support de la graphie. Malgré les avancées déjà réalisées auprès des différents organismes internationaux pour faire reconnaître la graphie et l'inclure dans le système UNICODE, il faut continuer à participer à des commissions et organisations internationales comme le W3C pour inclure l'amazighe dans tous les autres standards.

Communauté : il faut constituer un groupe de travail et de réflexion composé d'institutions d'associations, de personnes, etc. spécialistes de l'amazighe, sur le développement et la modernisation de la langue. Ceci, afin d'encourager la promotion de la langue, recenser toutes les ressources, les bonnes pratiques, ainsi que les technologies existantes (une sorte de référence) permettant d'aider à rapprocher les personnes travaillant dans le domaine. Il faut également représenter la langue amazighe au niveau des instances internationales pour aider à sa normalisation linguistique.

7. Perspectives et conclusion

Dans cet article, nous avons défini le LOD et rappelé quelques technologies le constituant. Nous avons mis l'accent sur certains aspects, qui nous ont permis de mettre en valeur, des problèmes dont souffrent les langues peu dotées ; comme dans le processus d'intégration des données (constitution de corpus, structuration, annotation, publication, etc.). Une partie de ces éléments a été recensée suite à une tentative de mise en place d'une ontologie scientifique en langue amazighe. Durant ce projet, plusieurs problèmes technologiques ou liés à la prise

en charge des contenus ont été rencontrés. Il y a par exemple ceux liés : à la création d'une terminologie scientifique par rapport au domaine étudié, à la prise en charge de la graphie Tifinagh, à l'absence d'outils de traitement, de normes linguistiques, etc. Suite à cela, nous avons passé en revue quelques recommandations pouvant aider à l'intégration de l'amazighe dans le LOD. Ceci afin d'enrichir, moderniser et préparer le patrimoine amazighe pour les nouveaux systèmes, gage de son développement et de sa pérennisation. Des éléments de ces recommandations constituent les perspectives du présent article.

Bibliographie

Auer S., Bizer C., Kobilarov G., Lehmann J., Ives. Dbpedia Z. (2007). A nucleus for a web of open data. *In 6th International Semantic Web Conference*, Busan, Korea, pp. 11-15. Springer.

Besacier L., Le V-B., Castelli E., Sam S., Protin L. (2005). Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer, Atelier « TALN et langues peu dotées », *TALN 05*, vol 2, pp. 207-217, Dourdan, France.

Le V-B., Bigi B., Besacier L., Castelli E. (2003). Using the Web for fast language model construction in minority languages, *8th European Conference on Speech Communication and Technology* (Eurospeech'03), pp. 3117-3120, Geneva, Switzerland.

http://inkdroid.org/lod-graph/

Vers un Dictionnaire Electronique de l'Amazighe

Fatima Zahra Nejme[1] Siham Boulaknadel[2] Driss Aboutajdine[1]
[1] LRIT, Unité Associée au CNRST (URAC 29),
Faculté des Sciences, Mohammed V-Agdal, Rabat, Maroc.
Fatimazahra.nejme@gmail.com
aboutaj@fsr.ac.ma
[2] IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
Boulaknadel@ircam.ma

Résumé

Depuis l'antiquité, le patrimoine amazighe est en expansion de génération en génération. Dans l'objectif de sauvegarder, exploiter ce patrimoine et éviter qu'il soit menacé de disparition, il semble opportun d'équiper cette langue de moyens nécessaires pour affronter les enjeux d'accès au domaine des nouvelles technologies de l'information et de la communication (NTIC). Dans ce contexte, et dans les perspectives de développer des outils et des ressources linguistiques pour le traitement automatique de la langue amazighe, nous avons entrepris d'utiliser la plateforme d'ingénierie linguistique NooJ afin de créer un module pour la langue amazighe standard (Ameur *et al*., 2004-a). Pour ce faire, étant donné que la construction d'un dictionnaire est une étape importante et fondamentale dans le traitement automatique d'une langue donnée, l'objectif principal de cet article est la construction d'un dictionnaire électronique pour la langue amazighe standard, et qui se basent en ce moment sur les deux catégories : nom et particules avec un ensemble d'informations morphologiques liées a chaque entrée.

1. Introduction

La langue amazighe du Maroc est considérée comme un constituant éminent de la culture marocaine et ce par sa richesse et son originalité. Cependant, il a été longtemps écarté sinon négligé en tant que source d'enrichissement culturel. Toutefois, au cours des dernières années, la société marocaine a connu beaucoup de débat sur la langue et la culture amazighe. Ainsi, la création d'une nouvelle institution gouvernementale, à savoir l'Institut Royal de la Culture Amazighe (IRCAM), a permis à cette langue aussi que sa culture à retrouver leur place légitime dans de nombreux domaines. Par conséquent, cette langue a pu être aménagée et son introduction assurée dans le domaine public notamment dans l'enseignement, l'administration et les médias. Cette création lui a permis d'avoir une graphie officielle, un codage propre dans le standard Unicode, des normes appropriées pour la disposition d'un clavier amazighe et des structures linguistiques qui sont en phase d'élaboration avec une démarche progressive. Cette démarche a été initiée par la construction des lexiques (Kamel, 2006; Ameur *et al*., 2009), l'homogénéisation de l'orthographe et la mise en place des règles de segmentation de la chaîne parlée (Ameur *et al*., 2006), et par l'élaboration des règles de grammaire (Boukhris *et al*., 2008). De ce fait elle a eu sa chance de se positionner dans la société globale de l'information.

Toutefois, en traitement automatique du langage naturel (NLP), l'amazighe, comme la plupart des langues non européennes[1], souffre encore de la rareté des outils de traitement automatique du langage ainsi que des ressources linguistiques, ce qu'elle ne permet pas à cette langue de rejoindre ses consœurs dans le domaine des nouvelles technologies de l'information et de la communication (NTIC). En ce sens, étant donné que la construction d'un dictionnaire est une étape importante et fondamentale dans le traitement automatique d'une langue donnée, l'objectif principal de cet article est la construction d'un dictionnaire électronique pour la langue amazighe standard du Maroc, et qui se basent en ce moment sur les deux catégories : nom et particules avec un ensemble d'informations morphologiques liées a chaque entrée. Pour ce faire, nous avons opté pour l'utilisation de la plateforme linguistique de développement NooJ, compte tenu de ses avantages, afin de construire un module pour l'amazighe, dont l'objectif est de l'utiliser dans l'enseignement au Maroc.

Le présent article se structure autour de trois volets: le premier présente un descriptif des particularités de la langue amazighe du Maroc, le deuxième expose le module NooJ ainsi qu'une description de notre dictionnaire, alors que le dernier volet est consacré à la conclusion et aux perspectives.

2. Particularités de la langue amazighe

2.1. Historique

La langue amazighe connue aussi sous le nom du berbère ou Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ), est une branche de la famille de langue afro-asiatique (chamito-sémitique) (Greenberg, 1966; Ouakrim, 1995) séparée en deux : langues berbères du Nord et du Sud. Elle présente la langue d'une population appelée « Imazighen » qui s'est installée depuis l'antiquité sur un espace géographique immense, et qui se présente à l'heure actuelle dans une dizaine de pays allant depuis le Maroc, avec 50% de la population globale (Boukous, 1995), jusqu'à l'Égypte, en passant par l'Algérie avec 25%, la Tunisie, la Mauritanie, la Libye le Niger et le Mali (Chaker, 2003). Au Maroc, l'amazighe se répartit selon deux types de dialectes: les dialectes régionaux et les dialectes locaux. Pour le premier type, il se répartie en trois grandes zones régionales qui couvrent l'ensemble des régions montagneuses: le Tarifit au Nord, le Tamazight au Maroc central et au Sud-Est et le Tashelhit au Sud-Ouest et dans le Haut-Atlas. Chacun de ces dialectes comprend des sous-dialectes ou dialectes locaux constituant le deuxième type. Ces sous-dialectes sont presque aussi nombreux que les régions où elles sont encore parlées. A titre d'exemple, le dialecte régional Tamazight contient un ensemble de sous-dialectes, dont nous citerons: le Tamazight de Béni-Mellal, le Tamazight d'Errachidia, le Tamazight de Ait Sadden, etc.

L'amazighe est une langue qui a tous ses attributs: son codage, son alphabet, sa grammaire, son orthographe et sa littérature orale extrêmement riche. Elle connaît une grande richesse au niveau de son vocabulaire. Ainsi, un seul sens est rendu de plusieurs façons dans chaque dialecte ou sous-dialecte. Par exemple : tête = « ixf, aqrru, ukhsas, azllif, axshash, ajdjif ».

1 Langues peu dotées informatiquement (les langues-À (Berment, 2004)).

Notre étude est focalisée, dans cet article, sur l'amazighe standard du Maroc. Depuis quelques années, le Maroc s'est engagé pour réaliser un processus de standardisation[2] de la langue amazighe (Ameur *et al*., 2004-a), qui a pour vocation d'uniformiser les structures et à atténuer les divergences, en éliminant les occurrences non distinctives qui entrainent souvent des problèmes d'intercompréhension. Ce processus de standardisation consiste à :

- Adopter une graphie standard normalisée sur une base phonologique ;
- Adopter un lexique de base commun ;
- Appliquer les mêmes règles orthographiques, les mêmes consignes pédagogiques, et les mêmes formes néologiques ;
- Exploiter la variation dialectale afin de sauvegarder la richesse de la langue.

2.2. Caractéristiques de l'amazighe standard du Maroc

Le but de cette section est de donner un aperçu sur les caractéristiques de la langue amazighe standard du Maroc qui incluent le système d'écriture graphique, le codage Unicode et les propriétés morphologiques des principales catégories syntaxiques.

2.2.1. Système d'écriture

En se basant sur le système original, l'IRCAM a développé un système d'alphabet sous le nom de Tifinaghe-IRCAM. Il s'écrit de gauche à droite. Cet alphabet standardisé est basé sur un système graphique à tendance phonologique. Cependant, il ne retient pas toutes les réalisations phonétiques produites, mais uniquement celles qui sont fonctionnelles (Ameur *et al*., 2004-b). Il est composé de 27 consonnes, 2 semi-consonnes, 3 voyelles pleines et une voyelle neutre.

2.2.2. Encodage Unicode

Depuis l'adaptation de Tifinaghe comme graphie officielle au Maroc pour la langue amazighe, l'encodage Tifinaghe est devenu nécessaire. Pour cette raison, des efforts considérables ont été investis par le centre d'études informatiques, systèmes d'information et de communication de l'IRCAM.

Le codage Unicode est constitué de quatre sous-ensembles de caractères Tifinaghe: l'ensemble de base de l'IRCAM, l'ensemble étendu de l'IRCAM, et d'autres lettres néo-Tifinaghe ainsi que des lettres Touareg moderne. Les deux premiers sous-ensembles constituent les ensembles de caractères choisis par l'IRCAM.

2 La standardisation de l'amazighe s impose d'autant plus avec son introduction dans le système éducatif, et avec le rôle que cette langue est appelée à jouer «dans l'espace social, culturel et médiatique, national, régional et local» (cf. article 2 du Dahir portant création de l'IRCAM).

2.2.3. Morphologie

La langue amazighe présente une morphologie riche. Elle peut être considérée comme une langue complexe dont les mots peuvent être classés en trois catégories morphosyntaxiques: Nom, Verbe et Particules (Boukhris *et al*., 2008).

a. Nom

En amazighe, le nom est une unité lexicale formée d'une racine et d'un schème. Il possède deux caractéristiques, la première est qu'il peut prendre différentes formes à savoir: une forme simple (ⴰⵔⴳⴰⵣ [argaz] "homme"), forme composée (ⴱⵓⴼⵢⵓⵀ [buhyyuf] "la famine") ou bien forme dérivée (ⴰⵎⵙⴰⵡⴰⴹ [amsawaḍ] "la communication"). La deuxième caractéristique correspond à la variation : il varie en genre (féminin, masculin), en nombre (singulier, pluriel) et en état (libre, annexion).

- ***Le genre***

Le nom amazighe connaît deux genres, le masculin et le féminin.

- Le nom masculin: il commence généralement par une des voyelles initiales: ⴰ [a], ⵉ [i] ou bien ⵓ [u], à titre d'exemple: ⵓⴷⵎ [udm] "visage", ⵉⵅⵀ [ixf] "tête". Cependant, il existe certains noms qui font l'exception: ⵉⵎⵎⴰ [imma] "(ma) mère".
- Le nom féminin: celui-ci est généralement de la forme ⵜ ....ⵜ [t....t], à l'exception de certains noms qui ne portent que le ⵜ [t] initial ou le ⵜ [t] final du morphème du féminin: ⵜⴰⵍⴰⵎⴰ [tadla] "gerbe", ⵕⵕⵎⵓⵢⵜ [rṛmuyt] "fatigue".

  Dans le cas général, le féminin est formé à partir du radical d'un nom masculin par l'ajout du morphème discontinue ⵜ ....ⵜ [t....t]: ⵉⵙⵍⵉ [isli] "marié" -> ⵜⵉⵙⵍⵉⵜ [tislit] "mariée".

- ***Le nombre***

Le nom amazighe, qu'il soit masculin ou féminin, possède un singulier et un pluriel. Ce dernier est obtenu selon quatre types: le pluriel externe, interne, mixte et le pluriel en id [id].

- Le pluriel externe: est obtenue par une alternance vocalique accompagné par une suffixation de n [n] ou l'une de ses variantes (ⵉⵏ [in], ⴰⵏ [an], ⴰⵢⵏ [ayn], ⵓⵏ [wn], ⴰⵓⵏ [awn], ⵓⴰⵏ [wan], ⵓⵉⵏ [win], ⵜⵏ [tn], ⵢⵉⵏ [yin]): ⴰⵅⵅⴰⵎ [axxam]-> ⵉⵅⵅⴰⵎⵏ [ixxamn] "maisons", ⵜⴰⵔⴱⴰⵜ [tarbat] -> ⵜⵉⵔⴱⴰⵜⵉⵏ [tirbatin] "filles".
- Le pluriel interne (ou brisé): est obtenue par une alternance vocalique plus un changement de voyelle internes (ⴰⵏⴰⵔ [adrar] -> ⵉⵏⴷⵓⵔ [idurar] "montagnes").
- Le pluriel mixte: est formé par une alternance d'une voyelle interne et/ou d'une consonne plus une suffixation par n [n] (ⵉⵍⵉ [ili] "part"-> ⵉⵍⴰ [ilan] "parts"); ou bien par une alternance vocalique initiale accompagné d'un changement vocalique final ⴰ [a] plus une alternance interne (ⴰⵎⵖⵖⴰⵔⵓ [amggaru] -> ⵉⵎⵖⵖⵓⵔⴰ [imggura] "derniers").
- Le pluriel en ⵉⵏ [id]: ce type de pluriel est obtenu par une préfixation de ⵉⵏ [id] du nom au singulier. Il est appliqué à un ensemble de cas de noms à savoir: des noms à

initiale consonantique, des noms propres, des noms de parenté, des noms composés, des numéraux, ainsi que pour les noms empruntés et intégrés (ⵅⴰⵍⵉ [xali] "(mon) oncle"-> ⵉⵏ ⵅⴰⵍⵉ [id xali]).

- ***L'état***

Nous distinguons deux états pour les noms Amazighs, l'état libre et l'état d'annexion.

- L'état libre: dans cet état, la voyelle initiale du nom ne subit aucune modification. Le nom est en état libre lorsqu'il s'agit: d'un mot isolé de tout contexte syntaxique, d'un complément d'objet direct, ou bien d'un complément de la particule prédictive ⵏ [d] "c'est".
- L'état d'annexion: cet état est fondé sur une modification de l'initiale du nom dans des contextes syntaxiques déterminés. Il prend l'une des formes suivantes: alternance vocalique ⴰ[a]/ⵓ[u] ou bien maintien de la voyelle initiale et ajout d'un ⵡ [w] au cas des noms masculins à initiale ⴰ [a] (ⴰⵔⵖⴰⵣ [argaz] "homme" -> ⵓⵔⵖⴰⵣ [urgaz]), addition d'un ⵡ [w] pour ceux à initial ⵓ [u] et d'un ⵢ [y] aux noms à voyelle ⵉ [i] (ⵉⵍⵉ [ils] "langue" -> ⵢⵉⵍⵉ [yils]). Pour les noms féminin, cet état est défini soit par la chute ou le maintien de la voyelle initiale (ⵜⴰⵎⵢⴰⵔⵜ [tamɣart] "femme"-> ⵜⵎⵢⴰⵔⵜ [tmɣart]).

b. Verbe

En amazighe, le verbe peut prendre deux formes : simple ou dérivée. Le verbe simple est composé d'une racine et d'un radical. Par contre le verbe dérivé est obtenu à partir des verbes simples par une préfixation de l'un des morphèmes suivants: ⵙ [s]/ss [ss], ⵜⵜ [tt] et ⵎ [m]/ ⵎⵎ [mm]. La première forme correspond à la forme factitive, la deuxième marque la forme passive et la troisième désigne la forme réciproque. Le verbe, qu'il soit simple ou dérivé, se conjugue selon quatre thèmes: l'aoriste, l'inaccompli, l'accompli positif et l'accompli négatif, et possède deux modes : l'impérative et l'impérative intensive.

c. Particules

Les particules sont un ensemble de mots Amazighs qui ne sont ni des noms, ni des verbes, et jouent un rôle d'indicateurs grammaticaux au sein d'une phrase. Cet ensemble est constitué de plusieurs éléments à savoir: les particules d'aspect, d'orientation et de négation; les pronoms indéfinis, démonstratifs, possessifs et interrogatifs; les pronoms personnels autonomes, affixes sujet, affixes d'objet direct et indirect, compléments du nom ordinaire et de parenté, compléments de prépositions; les adverbes de lieu, de temps, de quantité et de manière; les prépositions; les subordonnants et les conjonctions (Boukhris *et al*., 2008).

Généralement, les particules sont invariables. Or, dans le cas de l'amazighe, similairement au cas du français, il existe des particules flexionnelles telles que les pronoms possessifs (ⵡⵉⵏⵙ [winns] "le sien" -> ⵡⵉⵏⵙⵏ [winnsn] "le leur").

3. Traitement automatique de l'amazighe

3.1. La plateforme NooJ

NooJ (Silberztein, 2007) est une plateforme de développement linguistique qui offre un ensemble d'outils et méthodologies permettant de formaliser des langues tout en construisant, gérant et accumulant un grand nombre d'application de traitement automatique des langues (TAL), et les appliquant à des corpus de taille importante. Il permet de formaliser différents niveaux et composantes des langues naturelles, à savoir: l'orthographe, la morphologie (flexionnelle et dérivationnelle), le lexique (de mots simples, mots composés et expressions figées), la syntaxe locale et désambigüisation, la syntaxe, la sémantique et les ontologies. Pour chacun de ces niveaux, NooJ propose une méthodologie, un ou plusieurs formalismes adaptés, des outils-logiciels de développement et un ou plusieurs analyseurs automatiques de textes.

Actuellement, les utilisateurs de NooJ forment un public très varié en extension, ce qui a permis de développer des ressources linguistiques à large couverture dans une vingtaine de langues (arabe, arménien, bulgare, catalan, chinois, anglais, français, hébreu, hongrois, italien, polonais, portugais, espagnol, vietnamien et bélarussien).

Compte tenu de ces avantages, nous avons entrepris de construire un module NooJ pour la langue amazighe. L'objectif principal visé par ce travail est la construction d'un dictionnaire électronique contenant un ensemble d'informations morphologiques à savoir : le genre (masculin, féminin), le nombre (singulier, pluriel) et l'état (libre, annexion).

3.2. Représentation formelle du dictionnaire amazighe

Afin de développer un module NooJ pour l'amazighe standard du Maroc, nous avons entrepris la construction d'un ensemble de ressources linguistiques. Ainsi, nous avons commencé, dans cette contribution, par l'élaboration de notre dictionnaire électronique dédiés au traitement automatique de l'amazighe et qui a comme principale particularité d'associer toutes les entrées lexicales à un ensemble d'informations linguistiques. Notre première version du dictionnaire contient, à l'heure actuelle, 4488 entrées basé sur les deux catégories : noms et particules, et dont chaque entrée est non ambiguë. Nous présentons par la suite la structure des entrées de notre dictionnaire.

3.2.1. Structures des entrées

Chaque entrée dans le dictionnaire présente généralement les items suivants :

- Le lemme,
- La catégorie lexicale,
- La traduction en français,
- L'ensemble des informations morphologiques : genre, nombre et état (pour les noms).

3.2.2. Catégorie grammaticale

La première information que nous avons attribuée aux entrées de notre dictionnaire correspond à la catégorie grammaticale désignée par un code écrit en majuscules. Le tableau suivant présent les abréviations utilisées pour chaque partie du discours :

Catégorie	Code	Exemple
Non	N	"ⵄⵔⴳⴰⵣ" (argaz- homme)
Verbe	V	"ⵄⵡⵙ" (aws- aide)
Adjective	ADJ	"ⵓⴼⴼⵔ" (uffir- clandestin)
Adverbe	ADV	"ⴷⴰ" (da- ici)
Préposition	PREP	"ⵔ" (i- à)
Prenons	PRON	"ⵏⴽⴽ" (nkk- moi)
Démonstrative	DEM	"ⵄⵏ" (ann- là)
Relative	REL	"ⵎⵔ" (mi- qui)

Tableau 1 : Liste des catégories grammaticales

3.2.3. Trait syntactico-sémantiques

En deuxième lieu, nous avons introduit des champs d'informations syntactico-sémantiques aux entrées du dictionnaire. Ces informations sont citées dans le tableau ci-dessus.

Trait syntactico-sémantique	Code	Exemple
Abstrait	Abs	"ⵜⴰⵢⵔⵉ" (tayri- amour)
Animal	Anl	"ⵓⵛⵛⵍ" (uccn- chacal)
Concrète	Conc	"ⵜⴰⴷⵍⵉⵙⵜ" (tadlist- petit livre)
Humain	Hum	"ⵄⵎⵙⵡⵓⵔⵉ" (amswuri- operateur)
Alimentation	Alim	"ⵜⴰⵡⵙⴰⵢⵜ" (tavsayt- courgette)
Médical	Medic	"ⵜⵉⵙⵙⵏⵉⵜ" (tisgnit- seringue)
localisation	Loc	"ⵔⴱⴰⵜ" (rbat- Rabat)
Date/ heure	Date	"ⵢⵓⵏⵢⵓ" (yunyu- Juin)
Organisation	Org	"ⵜⴰⵎⵙⵙⵏⵜⵉⵜ" (tamssntit- entreprise)

Tableau 2 : Liste des traits syntactico-sémantiques

En outre ces informations linguistiques, un ensemble de caractéristiques morphologiques sont associées, à savoir: le genre, le pluriel et l'état.

3.2.4. Paradigme flexionnel

Compte tenu de la perspective purement informatique de notre dictionnaire, chaque entrée doit être rattachée à un ensemble d'informations morphologiques nécessaires pour que le système puisse générer et/ou reconnaître toutes les formes d'un mot.

Pour fléchir l'ensemble des entrées de notre dictionnaire nous avons dû créer, par le biais de graphes incorporé au logiciel NooJ, 331 graphes décrivant les modèles de flexions en amazighe et permettant de générer à partir d'une entrée lexicale ses informations flexionnelles : genre, nombre et état. L'une des principales caractéristiques de ces graphes réside dans le fait qu'elles opèrent au niveau du dictionnaire. En appliquant ces graphes nous obtenons un dictionnaire parallèle ayant toutes les formes fléchis regroupées par rapport à l'entrée de base. Ce regroupement est extrêmement utile au moment d'une analyse de texte pour faire des recherches soit par mot-forme soit par ensemble de formes fléchies correspondant à un lemme. Chaque entrée est associée à l'aide du code "+FLX" au modèle de flexion qui lui correspond. Nous montrons par la suite un exemple d'entrée de notre dictionnaire ainsi que la règle utilisée au niveau de sa flexion.

- ***Exemple :***

« ⴰⵍⵅⴰⵙ, N+FLX=A_C1+FR=Accident+Conc (angas- Accident) » : "ⴰⵍⵅⴰⵙ" est le lemme ; "N" est la catégorie lexicale, "+FLX=A_C1" est la règle de flexion à partir de laquelle l'ensemble des informations morphologiques vont être générer ; "+FR=Accident" : la traduction en français ; "+ Conc " est le trait sémantique.

- La règle morphologique "A_C1 " :

Cette règle de flexion correspond à la règle citée dans l'exemple précédent et qui permet de générer à partir d'un nom commençant par a [a] et se terminant par une consonne son correspondant féminin, son état d'annexion, son pluriel et le pluriel de son féminin.

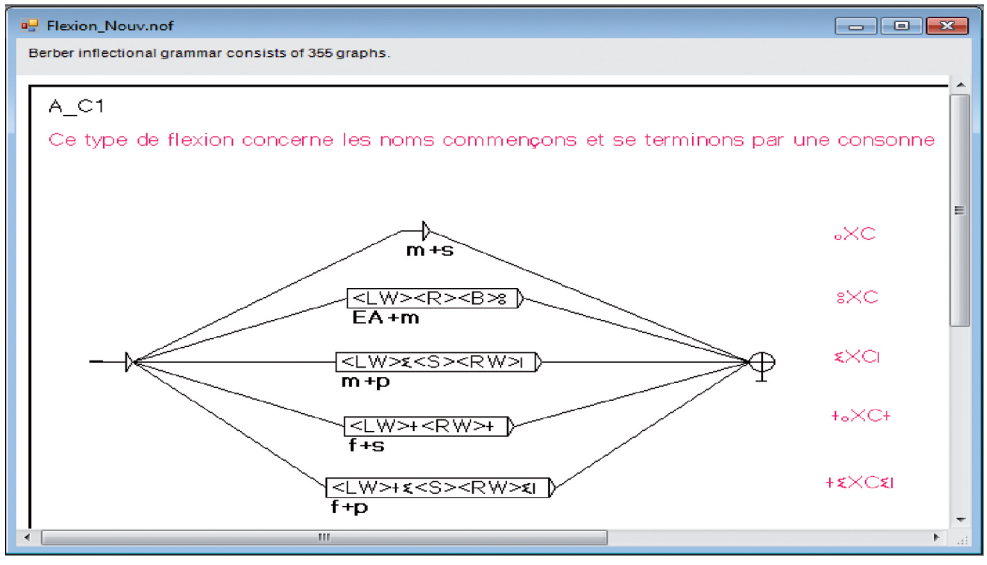


Figure 1 : Exemple de graphe avec le logiciel NooJ

Lorsque nous compilons notre dictionnaire électronique amazighe, le programme de flexion génère un total de 18,585 mots fléchis.

4. Conclusion et perspectives

Cet article décrit l'élaboration de notre dictionnaire électronique pour la langue amazighe standard en utilisant la plateforme linguistique de développement NooJ. Ce dictionnaire pourra être utilisé dans différentes applications en traitement automatique des langues en particuliers l'analyse des textes.

Dans la perspective d'améliorer ce travail, nous visons en premier lieu d'augmenter le nombre des entrées de notre dictionnaire ainsi que les informations morphologiques. Par ailleurs, il conviendrait d'ajouter l'autre catégorie verbe.

Références

Ameur M., Boumalk A. (2004). *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'Aménagement Linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut Royal de la Culture Amazighe, Série : Colloques et séminaires.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E., Souifi H. (2004). *Initiation à la langue amazighe*. Rabat, Maroc: IRCAM.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l'amazighe*. Rabat, Maroc : IRCAM.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009). *Vocabulaire des médias (Français-Amazighe-Anglais-Arabe)*. Série : Lexiques N°3, IRCAM, Rabat, Maroc.

Berment V. (2004). *Méthodes pour informatiser des langues et des groupes de langues peu dotées*, Thèse de doctorat de l'Université J. Fourier - Grenoble I, France.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*. Rabat, Maroc: IRCAM.

Boukous A. (1995). Société, langues et cultures au Maroc: Enjeux symboliques, Casablanca, Najah El Jadida.

Chaker S. (2003). Le berbère, Actes des langues de France, 215-227.

Greenberg J. (1966). *The Languages of Africa*. The Hague.

Kamel S. (2006). *Lexique Amazighe de géologie*. Rabat, Maroc: IRCAM.

Max S. (2007). *An Alternative Approach to Tagging*. NLDB 2007: 1-11.

Ouakrim O. (1995). *Fonética y fonología del Bereber*, Survey at the University of Autònoma de Barcelona.

Initiative pour le Développement d'un Corpus de la Langue Amazighe

Siham Boulaknadel **Fadoua Ataa Allah**
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Institut Royal de la Culture Amazighe, Rabat, Maroc
{boulaknadel, ataaallah}@ircam.ma

Résumé

Les corpus électroniques constituent de nos jours un élément essentiel et une base référentielle élémentaire, présentant les faits d'une langue, pour bien mener des recherches linguistiques, philologiques et informatiques. Conscientes de ce fait et dans la perspective de promouvoir la langue et la culture amazighes, nous avons opté au sein de l'Institut Royal de la Culture Amazighe de doter la langue amazighe d'un corpus de référence à visée exhaustive. Ainsi, nous avons entrepris l'élaboration d'un corpus constitué de textes de la langue amazighe permettant de mettre à la disposition de nos chercheurs une ressource représentant toutes les variantes et offrant une information en profondeur sur la langue amazighe.
Le présent article décrit les étapes entreprises au cours de la construction du corpus de la langue amazighe, et fait l'objet d'une évaluation sur la collection actuelle des textes.

1. Introduction

Depuis la création de l'Institut Royal de la Culture Amazighe (IRCAM), la langue amazighe au Maroc a bénéficié d'un statut institutionnel lui permettant d'avoir une graphie officielle, un codage propre dans le standard Unicode, des normes appropriées pour la disposition d'un clavier amazighe et des structures linguistiques qui sont en cours d'élaboration en empruntant une stratégie innovante et progressive. Cette stratégie a été initiée par la construction des lexiques (Ameur *et al*., 2006-b ; Kamel, 2006 ; Ameur *et al*., 2009), l'homogénéisation de l'orthographe et la mise en place des règles de segmentation de la chaîne parlée (Ameur *et al*., 2006-a) et par l'élaboration des règles de grammaire (Boukhris *et al*., 2008). Certes ces étapes de standardisation sont élémentaires et essentielles mais ne sont pas suffisantes pour qu'une langue peu dotée informatiquement telle que l'amazighe puisse franchir le seuil de la mondialisation informatique et de rejoindre ses consœurs dans ce domaine.

Dans cette perspective s'inscrit de nombreuses recherches scientifiques, principalement celles se focalisant sur la correction orthographique (Es-Saady *et al*., 2009), la traduction automatique (Rachidi et Mammas, 2007) la reconnaissance optique des caractères (Fakir *et al*., 2009), et celles s'occupant de la conception et la réalisation des ressources et outils linguistiques (Iazzi et Outahajala, 2008 ; Ataa Allah et Jaa, 2009 ; Boulaknadel, 2009 ; Ataa Allah et Boulaknadel, 2010-a; Ataa Allah et Boulaknadel, 2010-b; Outahajala *et al*., 2010 ; Boulaknadel et Ataa Allah, 2011). Or, pour mieux mener ce chantier de construction de ressources et outils linguistiques qui s'est ouvert à la langue amazighe, il s'avère primordial de doter la langue amazighe de corpus essentiel à son traitement automatique.

Dans ce travail, nous nous sommes intéressées à la construction d'un corpus dédié à la langue amazighe. Cette tâche s'inscrit dans le cadre d'un projet mené au sein de l'Institut Royal de la Culture Amazighe visant à fournir à la langue amazighe des ressources linguistiques riches et exploitables. Le but de cette initiative est d'encourager les linguistes à mener des recherches sur la langue amazighe, en créant une ressource utilisable pour cet objet. L'existence d'un tel rassemblement de textes en langue amazighe fournira également aux chercheurs intéressés par la langue amazighe un accès à des données actuellement dispersées ou non disponibles, et ce quel que soit leur domaine académique (linguistique, anthropologie, sociologie, littérature...). Grâce à la possibilité d'accès rapide aux données, et aux nouvelles possibilités de traitement de celles-ci, ce corpus informatisé permettra de nouveaux genres de recherches, auparavant non envisagés. Nous espérons que cette entreprise contribuera à donner une place à la langue amazighe dans la recherche linguistique du 21ème siècle.

Dans la suite de cet article, nous présentons dans la section 2 un descriptif des particularités de la langue amazighe standard du Maroc. Puis, nous détaillons dans la section 3 les étapes de construction de notre corpus ainsi qu'une batterie de mesures statistiques pour son analyse. Alors que nous consacrons la section 4 à la conclusion et aux perspectives.

2. Particularités de la langue amazighe

2.1. Historique

La langue amazighe connue aussi par le berbère est considérée comme la langue autochtone de l'Afrique du Nord (Hachid, 2000; Charles-André, 1978). Elle couvre toute l'Afrique du nord, le Sahara et une partie du Sahel ouest africain. Au Maroc, l'amazighe se répartit en trois grandes régions dialectales qui couvrent l'ensemble des régions montagneuses : au nord-est, le Rif avec le dialecte Tarifite ; au centre, le Moyen-Atlas et une partie du Haut-atlas avec le dialecte Tamazighte ; au sud et sud-ouest, le Haut-Atlas, l'Anti-Atlas et Souss, le domaine chleuh avec le dialecte Tachelhite.

Jusqu'à 1994 l'amazighe a été exclusivement réservée au domaine familial (Boukous, 1995). Cependant, suite au discours royal en 2001, l'amazighe est devenue une langue institutionnelle par la création de l'IRCAM. Et grâce à la constitution de 2011, l'amazighe a jouit auprès de sa consoeur l'arabe d'un statut d'une langue officielle.

2.2. Alphabet amazighe

L'amazighe fait partie des langues afro-asiatiques (Greenberg, 1966). Son système d'écriture, le « Libyco-berbère » (Tifinaghe en amazighe), date de plus de 25 siècles. Il est de nature alphabétique, à tendance phonologique fondé sur des signes à valeur consonantique, à usages traditionnellement assez restreints (funéraires, symboliques et ludiques). Cependant, les formats d'apparition de ses signes n'ont cessé de se développer : depuis son origine, le Libyque, jusqu'à le néo-tifinaghe, à la fin des années soixante, et le Tifinaghe-IRCAM, en 2001(Ameur *et al*., 2004). Ce dernier est orienté horizontalement de gauche à droite et composé de 27 consonnes, 2 semi-consonnes, 4 voyelles :

- 27 consonnes dont : les labiales (ⵒ, ⵕ, ⴼ), les dentales (ⵜ, ⵟ, ⴷ, ⴹ, ⵏ, ⵍ, ⵔ, ⵙ),

les alvéolaires (ⵟ, ⵙ, ⵚ, ⵣ), les palatales (ⵛ, ⵊ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⵯ, ⵅⵯ), les uvulaires (ⵇ, ⵅ, ⵖ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⵀ) ;

- 2 semi-consonnes : ⵢ et ⵡ ;
- 4 voyelles : trois voyelles pleines ⴰ, ⵉ, ⵓ et la voyelle neutre (ou schwa) ⴻ qui a un statut assez particulier en phonologie amazighe.

3. Construction et analyse du corpus

Malgré l'intérêt accordé à l'amazighe, il existe peu de travaux publiés concernant l'évaluation de corpus élaborés pour la langue amazighe standard du Maroc. Dans ce contexte, nous proposons à travers cet article de consacrer plus d'importance à ce genre de travaux, en partant de l'idée que tout traitement automatique de la langue amazighe ne peut se faire sans que cette dernière soit dotée d'un corpus de référence qui fera l'objet de recherches sur la langue. Ainsi, nous avons constitué et analysé un corpus de textes amazighes composé de 160 textes amazighes représentant différents genres littéraires (romans, poésie, contes, articles journalistiques) et couvrant différents thèmes.

3.1. Etapes de construction de corpus

L'élaboration de notre corpus se déroule en quatre étapes, dont la première consiste à collecter les documents écrits en langue amazighe, principalement ceux édités par l'IRCAM ou publiés dans son site officiel. La deuxième étape sert à la normalisation du format des documents collectés, où nous avons procédé à un dé-balisage des sources HTML et à une conversion des formats PDF et WORD en un format texte brut. Cette dernière sera succédée par une troisième étape qui consiste à convertir tous les textes en tifinaghe Unicode, en exploitant le convertisseur et translittérateur de la langue amazighe (Ataa allah et Boulaknadel, 2011). Par la suite, nous procédons à l'identification et la classification de chaque document selon sa thématique.

3.2. Propriétés statistiques du corpus

Dans le cadre de notre projet d'élaboration de corpus électronique pour la langue amazighe, nous avons collecté un corpus composé de 160 textes amazighes représentant différents genres littéraires, à savoir conte, conte pour enfants, poésie et articles de presse contenant les sous-genres journal, magasine et Net. Le Tableau 1 présente les caractéristiques statistiques du corpus collecté.

Statistiques	Conte	Conte pour enfants	Poésie	Net	Magasine	Journal	Global
Nombre de documents	4	21	10	78	11	36	160
Nombre de mots distincts	8,591	6,263	2,792	2,327	6,242	7,624	23,174
Nombre de mots total	37,363	29,741	8,534	9,519	23,086	27,850	136,093

Tableau 1 : Statistiques du corpus amazighe

Par ailleurs, nous avons eu recours à un ensemble de mesures statistiques afin d'évaluer et d'analyser notre corpus. Ainsi, nous nous sommes basées sur la loi Zipf-Mandelbort (Manning et Schütze, 1999), l'usage des caractères tifinaghes dans le corpus (Darrudi et Hejazi, 2004), les mesures de la richesse lexicale, et la représentativité des documents dans le corpus (Abdelali *et al*., 2005).

3.2.1. Loi de Zipf-Mandelbort

La loi de Zipf-Mandelbort est une distribution de probabilité discrète, connue également sous le nom de la loi de Pareto-Zipf (Zipf, 1949), qui est la forme continue de la loi de Zipf. Cette dernière prédit que si dans un texte de longueur N où les mots sont rangés dans l'ordre décroissant de leur fréquence d'apparition, la fréquence f(r) du mot de rang r est approximativement de forme $f(r) = k/r$, où k est une constante. Cette loi a été élargi par Mandelbrot en : $f(r) = \frac{A}{(B+r)^C}$, où A, B et C sont des constantes.

En utilisant l'outil Datafit pour l'ajustement des courbes, nous avons établi pour chaque catégorie de textes les graphes illustrés sur la Figure 1 qui présentent les occurrences des mots par rapport à leur rang.

D'après ces graphes, nous constatons que le comportement des mots de notre corpus représente bien la diversité et la nature du contenu de chaque catégorie de notre corpus. Généralement, la distribution de fréquence d'un corpus est séparée en 3 zones, à savoir :

- Zones à hautes fréquences dont la nature de ses mots sont essentiellement de type anti-dictionnaires. D'où l'inutilité d'étudier leur comportement vu que leur fréquence dépend mutuellement de la taille des documents du corpus.
- Zones à moyennes fréquences contenant, globalement, les termes représentant les différentes thématiques traitées par le corpus.

  Suite aux courbes de la Figure 1, nous constatons que pour les catégories conte, conte pour enfants et magasine la distribution de la fréquence des termes appartenant à cette zone est supérieure à celle de la loi de Zipf-Mandelbort. Tandis que nous remarquons l'inverse à l'égard des catégories poésie, journal et Net. Ce qui s'explique par le fait que les textes collectés pour les catégories poésie, journal et Net traitent plusieurs thématiques. Cependant, les catégories conte et conte pour enfants se composent principalement d'histoires destinées respectivement aux adultes et aux enfants. Ceux de la catégorie magasine introduisent les productions de l'IRCAM, en particulier les travaux en relation avec les contes et les contes pour enfants. Ce qui induit la mono-thématique de ces catégories.

  Par ailleurs, nous remarquons que la courbe du corpus global suit la même allure que celles des catégories conte, conte pour enfants et magasine. Ceci est dû au nombre des termes distincts de ces 3 dernières catégories qui représente 72.4% des termes du corpus global.

- Zones à basses fréquences dont la largeur dépend principalement de la variété du vocabulaire utilisé qui est liée à la qualité du style d'écriture.

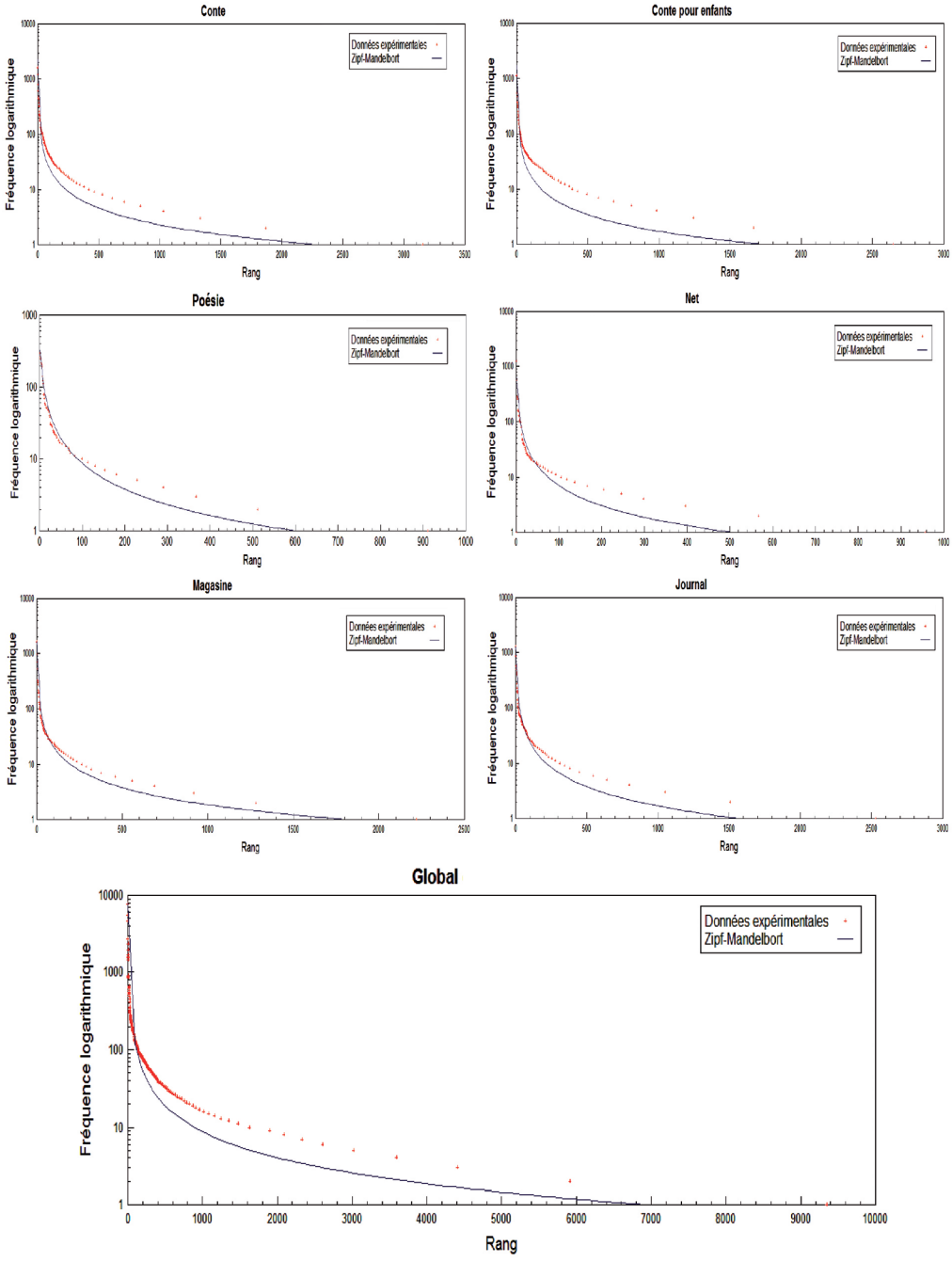


Figure 1 : Représentation schématique de la distribution des mots du corpus selon la loi de Zipf-Mandelbort

3.2.2. Usage des caractères tifinaghes

L'usage des caractères est une méthode d'évaluation qui consiste à calculer le pourcentage d'apparition de chaque lettre de l'alphabet de la langue étudiée dans le corpus élaboré, dans l'objectif de mesurer la richesse du corpus en terme de caractères. Dans ce contexte, nous avons procédé par le calcul de l'usage relatif des caractères tifinaghes, où la Figure 2 montre le pourcentage des occurrences de chaque lettre dans le corpus.

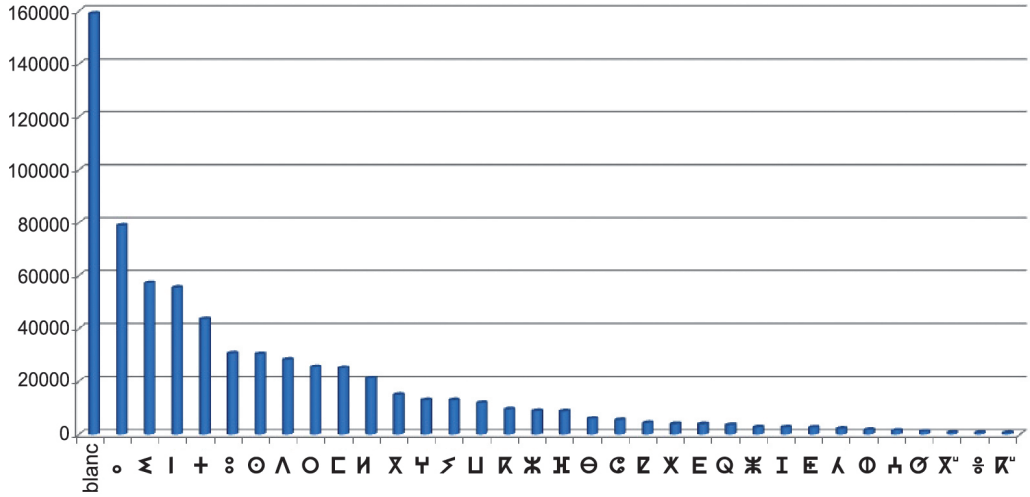


Figure 2 : Pourcentage des occurrences des caractères tifinaghes dans le corpus

D'après cette figure, nous pouvons remarquer que les caractères « ⵅⵓ » « ⵯ » et « ⴽⵓ » ne sont pas très utilisés dans le corpus par rapport aux autres caractères et que le caractère « espace » constitue 23% du corpus ce qui implique, à partir des occurrences des mots de notre corpus, que la moyenne de nombre de caractères par mot est à peu prés égale à 4,26.

3.2.3. Mesures de la richesse lexicale

La richesse lexicale est une notion intuitive et très subjective. Cependant, les mesures de la richesse lexicale cherchent à apporter une solution objective, mathématique, à un problème auquel les réponses n'ont été, pendant longtemps, que approximatives et impressionnistes. Nous allons dans ce qui suit présenter et commenter les résultats obtenus suite à l'application de deux méthodes de mesure de la richesse lexicale retenues.

i. TTR

Afin d'affiner nos analyses, nous procédons à une méthode de calcul de la richesse lexicale à partir d'un indicateur lexical, qui est le TTR ou Type Token Ratio :

TTR = V/N, où V est le nombre de mots distincts et N est le nombre total des mots ou l'étendue du texte.

Une lecture attentive des statistiques du Tableau 1 et du graphique de la Figure 3, représentant le classement des catégories de notre corpus selon la méthode TTR, nous montre une sensibilité de la richesse lexicale selon cette formule aux valeurs de l'étendue des textes.

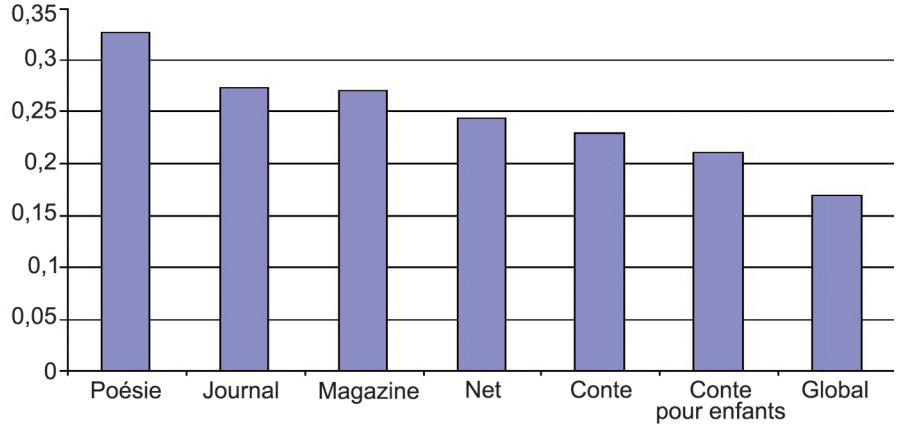


Figure 3 : Richesse lexicale selon la mesure TTR

ii. Indice W de Brunet

Dans le but de minimiser l'influence de l'étendue des textes sur la valeur de la richesse lexicale, Brunet (Brunet, 1978) fait jouer le rôle du facteur de réduction du nombre des mots distincts N à la réciproque du nombre total des mots V, une fois que ce dernier sera convenablement réduit à son tour par l'exposant fractionnaire α=0,172. La formule de l'indice W de Brunet s'écrit donc ainsi :

$$W = N^{V-a}$$

	N	V	V^{α}	$W = N^{V-\alpha}$
Global	136,093	23,174	5,63	8,15
Journal	27,850	7,624	4,65	9,02
Conte	37,363	8,591	4,75	9,18
Magazine	23,086	6,242	4,5	9,34
Conte pour enfants	29,741	6,263	4,5	9,87
Poésie	8,534	2,792	3,91	10,1
Net	9,519	2,327	3,79	11,18

Tableau 2 : Richesse lexicale selon l'indice W de Brunet

En analysant les résultats de cette formule représentés dans le Tableau 2, nous constatons que la catégorie journal qui se trouve en tête du classement des différentes catégories de notre corpus est la plus riche lexicalement, ce qui est dû principalement à la variété des sujets traités par cette dernière.

3.2.4. Représentativité des documents dans le corpus

La représentativité des documents dans le corpus permet d'évaluer l'apport de chaque document en terme de nouveaux mots ajoutés à la collection. Ainsi, nous avons entrepris cette évaluation par l'intégration d'un document par document à l'ensemble traité et la vérification de sa contribution à la construction de notre corpus, en calculant le nombre de mots distincts ajoutés suite à l'adjonction de ce document.

Nombre de documents	Nombre de mots	Nombre de mots distincts
10	29,693	6,930
20	39,421	8,995
40	44,761	9,986
80	88,119	16,824
120	100,424	18,539
160	136,093	23,174

Tableau 3 : Contribution des documents dans la richesse du corpus

En comparant le nombre des mots distincts pour chaque ensemble de documents représentés par le Tableau 3, nous constatons que la quantité de données ajoutées à chaque reprise contribue de manière significative à l'enrichissement du vocabulaire du corpus. Ainsi, nous pouvons conclure qu'à ce stade notre corpus n'a pas encore représenté toutes les variétés lexicales de la langue amazighe.

4. Conclusion

Cet article s'inscrit dans une démarche fondatrice d'élaboration de ressources et d'outils de traitement automatique de la langue amazighe, qui contribue à la promotion et le développement de cette langue. A ce titre, nous avons élaboré un corpus de textes à visée exhaustive pour la langue amazighe, que nous avons analysé et évalué en exploitant une batterie de mesures. A la base de ces mesures, nous retenons que ce corpus nécessite un enrichissement d'ordre horizontal et vertical. Dans le sens d'augmenter respectivement la richesse lexicale des catégories existantes et la variété thématique de notre corpus.

Références

Abdelali A., Cowie J., Soliman H. S. (2005). Building a modern standard Arabic corpus. *Actes du computational modeling of lexical acquisition workshop,* pp. 1-7.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l'amazighe*. Maroc : IRCAM.

Ameur M., Bouhjar A., Elmedlaoui M., Iazzi E. (2006). *Vocabulaire de la langue amazighe (français-amazighe)*. Maroc : IRCAM.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009). *Vocabulaire de la langue amazighe* (amazighe-arabe). Maroc : IRCAM.

Ataa Allah F., Jaa H. (2009). Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue Amazighe. Actes du 1er symposium international sur le traitement automatique de la culture amazighe. pp. 110-119.

Ataa Allah F., Boulaknadel S. (2010). Pseudo-racinisation de la langue amazighe. *Actes du Traitement Automatiques des Langues Naturelles*.

Ataa Allah F., Boulaknadel S. (2010). Online Amazigh Concordancer. *Proceedings of International Symposium on Image Video Communications and Mobile Networks*. Rabat, Maroc.

Ataa Allah F., Boulaknadel S. (2011). Convertisseur pour la langue amazighe : script arabe - latin - tifinaghe. Actes du *2ème symposium international sur le traitement automatique de la culture amazighe*. Agadir, Maroc, pp. 3-10.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*. Rabat, Maroc : IRCAM.

Boukous A. (1995), *Société, langues et cultures au Maroc : Enjeux symboliques,* Casablanca, Najah El Jadida.

Boulaknadel S. (2009). Amazigh ConCorde: an appropriate concordance for Amazigh. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 176-182.

Boulaknadel S., Ataa Allah F. (2011). Building a standard Amazigh corpus. *Proceedings of the International Conference on Intelligent Human Computer Interaction*. Prague, Tchec.

Brunet E. 1978. *Le vocabulaire de Jean Giraudoux. Structure et évolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la Langue Française*. Genève : Slatkine.

Charles-André J. (1978). *Histoire de l'Afrique du nord des origines à la conquête arabe: Tunisie - Algérie – Maroc*. France : Editions Payot.

Darrudi E., Hejazi M.R. (2004). Assessment of a Modern Farsi Corpus. Actes de *2nd Workshop on Information Technology & its Disciplines*.

Es Saady Y., Ait Ouguengay Y., Rachidi A., El Yassa M., Mammass D. (2009). Adaptation d'un correcteur orthographique existant à la langue Amazighe : cas du correcteur Hunspell. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 149-158.

Fakir M., Bouikhalene B., Moro K. (2009). Skeletonization methods evaluation for the recognition of printed tifinaghe characters. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 33-47.

Hachid M. (2000), *Les premiers berbères: entre méditerranée, Tassili et Nil*. France : Edisud.

Greenberg J. (1966). *The Languages of Africa*. Mouton, USA: The Hague.

Iazzi E., Outahajala M. (2008). Amazigh Data Base. *Actes de l'atelier HLT & NLP within the Arabic world: Arabic language and local languages processing status updates and prospects*. pp. 36-39.

Kamel S. (2006). *Lexique Amazighe de géologie*. Maroc : IRCAM.

Manning C., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. USA: MIT Press.

Outahajala M., Zenkouar L., Rosso P., Martí M. (2010). Tagging Amazigh with AncoraPipe. *Actes de Semitic Languages Workshop, 7th International Conference on Language Resources and Evaluation*. pp. 52-56.

Zipf G. K. (1949). *Human behaviour and the principal of least-effort*. USA: Addison-Wesley.

TICAM'12, p. 75-87

De la Taxinomie au Traitement Automatique des Verbes en Amazighe

L'houssaine El Gholb
Université Med V-Agdal.
elgholb@gmail.com

Résumé

Cet article tente d'aborder les problèmes linguistiques de la taxinomie des monèmes verbaux en amazighe. Pour ce faire, nous partons d'un essai de classification d'une base de données d'environ 4700 monèmes verbaux, à partir de l'examen de diverses habiletés et attitudes morphologiques qui les caractérisent. Notre objectif principal est d'inventorier les processus morphologiques investis par les classes verbales, lors de la formation de différents thèmes verbaux, pour une exploitation en Traitement Automatique des Langues (TAL). Devant l'étendue du domaine du TAL et l'hétérogénéité qui caractérise la morphologie verbale, le présent article essayera d'aborder le volet linguistique principalement, les propriétés morphologiques du verbe simple puisqu'il se situe en haut de la taxinomie des verbes et constitue la base de départ de la dérivation verbale.

1. Introduction

Le TAL a pour objet la création de programmes informatiques capables de traiter ou d'analyser automatiquement des langues naturelles. Cette tâche nécessite des compétences diverses : *i*) la compétence linguistique relative à la langue ; *ii*) la description formelle des connaissances linguistiques dans un formalisme qui convient au TAL ; *iii*) la compétence informatique pour créer des programmes informatiques capables de transformer un objet d'entrée en un objet de sortie (*cf*. Cardey et Mandic, 2005).

L'informatisation de l'amazighe est d'autant plus sollicitée avec le passage de l'oral à l'écrit et la volonté d'instaurer une langue qui peut relever le défi scientifique et littéraire. Sans vouloir intervenir sur la relation contingente entre la taxinomie et le TAL, notre article s'intéresse principalement aux problèmes linguistiques qui se posent pour la taxinomie des monèmes verbaux et qui peuvent entraver, effectivement, leur traitement automatique.

L'approche taxinomique est employée pour désigner la classification de suites d'éléments formant des listes définies, de façon précise et univoque, selon des traits distinctifs. En d'autres mots, la taxinomie sert à classer les monèmes, suivant les traits partagés, dans des classes

et des catégories[1]. Cela suppose des méthodes du repérage des relations morphologiques communes aux monèmes constituant ces classes et groupes. Néanmoins, la classification et l'organisation des monèmes verbaux se heurtent à des contraintes morphologiques diverses.

2. Les caractéristiques morphologiques du monème verbal

2.1. Au niveau dérivationnel

Les travaux réalisés sur la morphologie dérivationnelle[2] cantonnent la dérivation des synthèmes verbaux (causatif, réciproque, passif) au processus d'affixation et de combinaison de morphèmes dérivationnels. Certes, la forme qui sert de base pour la formation des dérivées et des sur-dérivées est la forme simple, dite « neutre » ou « nue », du monème verbal. Vraisemblablement, la forme dérivée résulte de la combinaison d'un morphème dérivationnel et la base verbale. Cette formation des dérivés se fait par la préfixation, par l'adjonction d'un morphème dérivationnel ou la combinaison de deux morphèmes de valeurs différentes[3].

2.2. Au niveau flexionnel

Le monème verbal, simple ou dérivée, reçoit des affixes qui expriment les informations de PNG (personne, nombre, genre) et d'aspect, lors de la conjugaison. Ces informations sont régies par les rapports grammaticaux qu'entretiennent les monèmes grammaticaux et lexicaux. Au niveau du paradigme de conjugaison, le monème verbal peut accepter quatre thèmes aspeco-temporels (aoriste, accompli, accompli négatif et inaccompli).

Sans prendre en considération les cas de syncrétisme qui neutralisent l'opposition thématique, un ensemble de procédés sont employés, selon le type ou la structure verbale, pour former les thèmes verbaux. D'ailleurs, la formation de l'accompli[4] se fait, en l'opposant au thème de l'aoriste, par une alternance ou une insertion vocalique. Par contre, l'inaccompli se forme, à la base d'une forme « nue », par préfixation du morphème « tt-» ou par gémination d'une consonne radicale. Ces deux procédés peuvent être employés seuls ou accompagnés d'une insertion vocalique, selon le type ou la structure verbale.

---

1 Le caractère équivoque de ces deux termes qui se confondent souvent, impose de leur donner une définition. Ainsi, Dubois *et al*. précisent, dans son dictionnaire, qu' « *une classe représente un ensemble d'unités linguistiques ayant une ou plusieurs propriétés communes entre elles* » (Dubois *et al*., 1994 : 86). Quant au terme de catégorie, Dubois (*ibid*) signale, à la page 78, qu'elle « *désigne une classe dont les membres figurent dans les mêmes environnements syntaxiques et entretiennent entre eux des relations particulières* ». Pour Martinet, une classe de monèmes «*présentent les mêmes compatibilités, à la condition qu'ils s'excluent mutuellement à un même point de chaîne*» (Martinet, 1985 : 108).

2 La morphologie dérivationnelle étudie la construction des monèmes lexicaux à partir de la combinaison de différents monèmes. Ainsi, le monème verbal simple, comme base de départ pour former des monèmes verbaux dérivés ou sur-dérivés, se compose d'une racine/radical et d'un schème.

3 Certainement, ces verbes dérivés présentent avec leur correspondant simple une relation de dépendance qui s'affiche aux niveaux morphologique et sémantique. Ils peuvent exprimer la valeur du factitif avec le morphème « s » ; la valeur du passif avec le morphème « ttu »; la valeur du réciproque et passif avec le morphème « m » et « n ».

4 Sauf syncrétisme, l'opposition accompli /accompli négatif est marquée morphologiquement par l'insertion de la voyelle « i » en position finale ou post-finale selon la structure verbale.

3. La taxinomie du monème verbal

3.1. Les méthodes classificatoires des amazighisants

Les comparatistes se sont penchés sur ce problème de la classification dès leurs premières études en prenant en considération les variations que présentent le verbe aux différents thèmes verbaux. Généralement, les méthodes de classification classiques reposent sur la notion de « classe du verbe ». En essence, le groupe/classe serait l'ensemble de verbes ayant les mêmes procédés d'affixation et/ou une même structure consonantique et un paradigme d'alternances vocaliques. Ainsi, la détermination de groupes ou classes peut se faire suivant: *i)* le nombre de consonnes de la base verbale ; *ii)* la qualité des consonnes (tendue, relâchée) ; *iii)* la nature des voyelles (i, u, a) ; *iv)* la position de la voyelle (initiale, médiane ou finale) ; *v)* l'alternance vocalique ; *vi)* l'alternance consonantique ; *vii)* la combinaison de ces deux derniers procédés (vocalique et consonantique).

Ces critères de classification varient d'un auteur à l'autre. En fait, Basset adopte une démarche stricte et une analyse rigoureusement détaillée en se référant aux particularités structurelles des monèmes verbaux (Basset, 1929). Dans ce sens, il distingue les verbes selon leur nombre de phonèmes consonantiques et leur comportement dans plusieurs parlers amazighes. Il en résulte une abondance de classes scindées au nombre de huit classes majeures, contenant largement des sous-catégories[5].

Par la suite, en adoptant la même classification initiée par Basset (*ibid.*), Dallet a distingué sept classes, pour le verbe en Kabyle (Dallet, 1953). Par contre, Laoust a établi, pour le parler de Ntifa, six classes verbales[6] en fonction de la nature de la voyelle, sa position (initiale, médiane et finale) et le type d'alternance vocalique (Laoust, 1918). Subséquemment, les travaux ultérieurs (Chaker, 1984 ; Cadi, 1987; Bentolila, 1981; entre autres) ont adopté la même démarche. Par contre, Aspinion a réduit, en jouant sur les interactions des classes verbales, le nombre en quatre groupes verbaux[7] basés sur la nature du phonème consonantique et vocalique (Aspinion, 1953).

5 Les monèmes à voyelles zéro et à première radicale brève, les monèmes à première radicale alternante, les monèmes à voyelle pleine, à alternance (pré-radicale, post-radicale, et intra-radicale), les monèmes à première radicale longue et les monèmes de qualité et verbe à voyelle alternante devant la dernière radicale.

6 La classe des verbes de type « $c^1c^2c^3$ » qui est invariable à l'accompli (exp*.skr* « faire »), la classe des verbes de type « c^1c^2 » qui a la forme « c^1c^2i/a » à l'accompli (*nġ*/ *nġi*/a), la classe des verbes de type « ac^1c^2 » qui a la forme « uc^1c^2 » à l'accompli (*amz*/*umz* « saisir »), la classe des verbes de type « c^1uc^2 » qui a la forme « c^1uc^2 » à l'accompli (*lal*/*lul* « naître »), la classe des verbes de type « c^1c^1u » qui a la forme « c^1c^2i/a » à l'accompli et la classe des verbes de type « $c^1c^2ic^3c^4$ » qui a la forme « $c^1c^2ac^3c^4$ ».

7 Le groupe à thème constant, le groupe à variation vocalique simple, le groupe à double variation vocalique et le groupe à variation mixte (vocalique et consonantique).

3.2. Méthodes de classement traditionnelles

Les méthodes traditionnelles proposées sont basées sur la structure verbale suivant les critères énoncés auparavant (type de monème, schème, etc.). Généralement, deux classes sont à distinguer d'un côté, les verbes réguliers qui présentent un radical verbal constant et une même structure morphologique lors de la formation des thèmes verbaux (syncrétisme d'aoriste/prétérit, par exemple) ; de l'autre, les verbes qualifiés d'irréguliers qui modifient leur structure d'un thème à l'autre (forme différente aoriste/prétérit, par exemple).

Cette distinction s'opère aussi à la base des monèmes simples et composés (dérivés et sur-dérivés). Par ailleurs, d'autres classifications sont basées principalement sur les variations structurelles qui dépendent de comportement morphologique des monèmes verbaux. Par conséquent, les monèmes verbaux peuvent être purement consonantiques (consonne simple ou géminée) et/ou à voyelle, à l'intérieur desquels sont définies des sous-classes suivant le nombre et le caractère tendu ou relâché des phonèmes consonantiques, ou suivant le timbre et le caractère constant ou non constant, alternant ou non alternant des voyelles.

Par contre, la classification adoptée par les structuralistes est basée sur le nombre des thèmes (aoriste, prétérit, prétérit négatif, inaccompli) qu'un verbe peut admettre. Ainsi, ils distinguent trois classes des verbes (les verbes à quatre, à trois, à deux thèmes). Ces classes emboîtent plusieurs sous-classes suivant la nature des thèmes identiques et le type de variations mises en jeu (*cf.* Boukhris 1986, pour une classification systématique des différents types).

4. Apports et limites des approches de classification

Nous allons essayer d'évaluer les procédés classificatoires adoptés dans les taxinomies précédentes. *A priori*, les amazighisants ne sont pas toujours d'accord sur la méthode de classer les monèmes verbaux; la liste reste ouverte, aussi n'est-elle pas toujours la même. Dans cette perspective, il n'existe pas de méthode moins coûteuse pour l'amazighe. Toutes les classifications proposées ont une part de vérité quelque part et causent de sérieux problèmes pour une organisation plus logique. Les classements sont si nombreux que nous nous demandons quelle serait leur utilité (*cf.* le Manuel de conjugaison du CAL, 2012). Par ailleurs, il est à signaler qu'il n'existe pas de classification des verbes en catégorie comme en français, par exemple.

Pour présenter les monèmes verbaux dans des groupes assez homogènes en tenant compte de différents types de particularités qui les unissent, les amazighisants proposent des classifications axées sur les phonèmes constitutifs, leur agencement, l'opposition thématique et le nombre des thèmes pour chaque verbe, entre autres. Ces classifications n'ont pas réussi, certainement, parce qu'elles présentent des défiances disparates.

D'ailleurs, l'approche basée sur l'analogie de la structure morphologique reste contestable étant donné que les monèmes verbaux, même s'ils ont une même structure, ils peuvent avoir deux comportements morphologiques différents (exemple, *af* « trouver », « être meilleur » qui donne respectivement *yufa* par opposition à *yuf*). Il est à signer que les données erronées obtenues découlent, *a priori*, des données de départ non fondées. Par conséquent, de l'abondance des classes et la divergence de la conjugaison résulte une différence du classement.

Le maillon faible dans ces classifications concerne, premièrement, le choix de la forme neutre (forme de base)[8] qui se présente sous diverses formes concurrentes avec des conjugaisons distinctes (*cf.* Iazzi, 1991). A titre d'exemple, le verbe *ẓr* et *iẓir* « voir », *ssnetisin* « connaître », etc., qui présente au moins deux variantes régionales pour la forme d'aoriste. Ce qui engendre une différence au niveau des procédés de conjugaison (exemple : *ttidir* et *zrra* à l'inaccompli)[9].

En plus de ces variations morphologiques, il apparait que l'impact des processus phonologiques, surtout de l'assimilation est considérable. Par exemple, les deux formes de l'inaccompli: *nkkr* et *kkar* sont deux allomorphes du même verbe *nkr* « se réveiller ». Ces deux réalisations prolifèrent une anarchie au niveau de la morphologie verbale et bloquent toute entreprise de modélisation de la conjugaison amazighe. Généralement, le phénomène d'assimilation peut être à l'origine de cette défiance que connaissent la conjugaison amazighe et la cause de glissement des verbes d'un groupe à un autre.

Également, les radicales faibles résultant de la vocalisation des semi-consonnes « *w* » et « *y* », respectivement du « u » et « i » sont responsables de la déficience de la conjugaison. Ce changement structurel n'est cependant pas facile à analyser, et c'est à juste titre que les semi-consonnes sont considérées, selon Basset (*idid*.), comme « le poison du berbère ». Selon Bentolila, l'apparition de « *y/i* -» dépend de l'action du thème « *avec telle ou telle modalité personnelle sujet* » (Bentolila, 1981). Néanmoins, rien n'est bien établi quant à leur vocalisation. Il y a des cas où elles se maintiennent même en finale du verbe et dans le contexte approprié. En principe, ce jeu d'alternances est dû à la phono-tactique de la langue (*cf.* Boukous, 1987).

Dans un autre lieu, la réduction de la racine peut affecter l'organisation des verbes en classes soit au niveau de la structure ou de la conjugaison. Ainsi, la base verbale peut se présenter sous une forme amputée à cause d'un amenuisement de la racine10. La soustraction pose un problème, même s'il ne gêne pas la reconnaissance des lexèmes verbaux, quand il s'agit du classement par structure et le choix de la classe appropriée11. Par exemple, nous avons toute une série de formes pour la racine signifiant « donner » : au centre : fk, kf, k (tikkit : nom d'action verbale), sans oublier les formes rétrécies du Nord : uš, où le phonème « f », attesté comme verbe au Sud, est tombé et où le phonème « k » est passé à « š ».

8 L'idée générale est de postuler une forme unique pour chaque lexème verbal qui débouchera sur des règles morphologiques généralisables sur les autres variantes. Il convient de « *choisir les formes de base qui permettront la description la plus simple* » (Gleason, 1969).

9 Nous pourrions facilement multiplier les exemples, *irid* et *arud*' « laver », *inig* et *anug* « chercher », *ismum* et *smim* « rendre acide », mais il en est parfois autrement pour : *kkis /kks* « enlever », *uš/ fk/ kf/ḳ* « donner », *zzg/zzi* « traire », *warg /arji* « rêver », *rrz/rz* « casser », l'inac.).

10 La racine est l'ensemble de consonnes qui reçoit sa coloration phonique par l'introduction de voyelles et d'affixes et façonne de ce fait sa configuration formelle et sémantique (*cf.* Cantineau, 1950).

11 Le raisonnement qui nous paraît adéquat est qu'effectivement, ces verbes sont le résultat d'une « évolution » qui a comme conséquence « *la désagrégation du système* » ou bien qu'ils ont des processus particuliers encore non déterminés de manière satisfaisante. Lionel Galand parle d'une certaine « *usure du système* » à propos d'un phénomène similaire dans la morphologie nominale (Galand, 1977).

Par ailleurs, certaines formes verbales connaissent le phénomène de la « suppléance ». Ce problème concerne les changements et les irrégularités observés dans certains thèmes supplétifs. Ainsi, la formation de l'inaccompli n'est pas constante pour les verbes qui changent du radical, en plus de jeu d'affixes (suffixes ou préfixes) pour indiquer sa valeur aspecto-temporelle. En effet, la conjugaison de ces verbes fait partie des connaissances du locuteur. Considérons le réseau de relations de dépendance entre les différents thèmes de verbe *ini* « dire » qui devient *nni/a* à l'accompli et *ttini* à l'inaccompli ; en l'occurrence, *qqar* au Nord. Ensuite, les verbes relevés dans le parler des Ait Seghrouchens : *isi/ittsi*, asy/ttasy « prendre » et *uss* « taire » qui se conjugue au prétérit avec le radical *susm*, *ssusm*. Ce type de verbe irrégulier change partiellement du radical d'un thème à l'autre. Par conséquent, l'exception morphologique rend complexe la conjugaison de ces formes supplétives.

5. Vers le traitement automatique des verbes

Le TAL, tel que défini dans notre introduction, fait intervenir des recherches diverses dans différents domaines pour effectivement traiter les données linguistiques. Dans une première approximation, le traitement d'une langue concerne la transformation d'un objet d'entrée en un objet de sortie[12] sous forme de données brutes. Cette transformation comprend une étape intermédiaire qui vise à extraire les informations linguistiques (*cf.* Xanthos, 2008).

Lors de l'inventaire des informations morphologiques, il est avéré que ces informations se subdivisent en connaissances flexionnelles, dérivationnelles (différentes formes flexionnelles, dérivationnelles). Ces connaissances morphologiques présentent un ensemble de données et de règles qui permettent aux systèmes d'identifier des monèmes verbaux en leur associant toutes les données pertinentes pour la suite du traitement.

Avant de se pencher sur les types de relations qu'entretiennent le monème verbal et les informations morphologiques, il est important de faire apparaître leur complexité et les différentes organisations. Autrement dit, l'établissement des relations entre le monème verbal et toutes les formes flexionnelles ou dérivationnelles qui lui sont associées morphologiquement. Ainsi, la forme de base (verbe simple) est liée à ses flexions et ses formes dérivées.

Le monème verbal est généralement construit à partir de plusieurs éléments. Il est constitué d'une partie centrale, appelée racine. Les connaissances morphologiques concernent la manière dont les mots sont construits à partir des unités minimales de signification (monèmes/morphèmes, les affixes, etc.).

12 Nous parlons d' « analyse » lorsque le point d'arrivée ou l'output sont des données brutes et de la « génération » lorsque le point de départ ou l'input sont des données brutes.

5.1. Problème de classification

Les monèmes verbaux sont classés, d'un point de vue morphologique, selon leurs phonèmes consonantiques constitutifs. Nous aurons, ainsi, des verbes monolitères constitués d'un phonème consonantique, les bilitères constitués de deux phonèmes consonantiques, les trilitères formés de trois phonèmes consonantiques. Certes, ces verbes sont formés à partir des bases simples avec/sans voyelle. Ces bases se composent aussi selon la gémination ou non du radical, de la base avec/sans gémination.

Ces classements diffèrent sur plusieurs points, et nous pouvons nous interroger non seulement sur le rendement pédagogique d'une telle méthode mais aussi sur l'efficacité de chaque méthode. C'est quoi la valeur pratique de ces classements ? Certes, ces classements servent pratiquement pour présenter les verbes en des classes homogènes et régulières pour faciliter, à la fois, leur traitement informatique et leur enseignement/acquisition à l'école.

Les problèmes posés par la taxinomie et la classification des verbes se manifestent aux différents niveaux et se trouvent, *a priori*, dans la définition des classes et des groupes des verbes qui se heurtent à des contraintes diverses. Le problème d'irrégularité et d'homogénéité des classes sont rendus plus épineux encore lorsqu'il s'agit d'une exploitation liée au TAL. En effet, la plupart des verbes connaissent des variations apparentes, sur lesquels nous ne nous attarderons pas, car étudier tous ces variations irait bien au-delà de notre modeste article.

Il s'avère que la majorité des études s'efforcent de saisir les processus de la morphologie verbale strictement au niveau de la surface. En occultant des phénomènes phonologiques et phono-tactiques relatifs à la langue, cette position réduit les modifications qui relèvent de la dérivation et la formation morphologique en tant qu'opération morphologique. Néanmoins, les taxinomies proposées dans ces travaux reposent sur des critères hétéroclites. Pour certains, le critère de classement se base sur la forme ou la structure du monème. Pour d'autres, la nature du changement observé pour la formation des thèmes verbaux sert de principe classificatoire.

Au vue des travaux classiques de classification, des efforts ont été apportés par les successeurs pour réduire et harmoniser les classes verbales. Ces efforts ont pris en considération les soucis didactiques de l'enseignement de la morphologie amazighe. Compte tenu de l'importance des données, le travail est donc de trouver des verbes types ou similaires qui rendent compte des autres verbes d'une même classe.

Généralement, une classe est définie par référence à un ensemble de propriétés qui sont à la fois nécessaires et suffisantes pour appartenir à une classe. Autrement dit, la classification est un processus du regroupement des monèmes verbaux en une hiérarchie de classes suivant certaines propriétés communes. Par ailleurs, la combinaison des procédés de formation peuvent s'interagir entre eux et amplifier d'avantage les classes verbales.

5.2. Schémas classificatoires des monèmes verbaux

5.2.1. Schéma basé sur la morphologie dérivationnelle

Les verbes simples, comme nous l'avons déjà précisé, se situent tout en haut de la taxinomie des verbes : cela montre qu'il peut être employé comme base de dérivation des autres verbes et que ses propriétés seront par conséquent moins restreintes. De plus, ce type de verbes est le plus fréquent dans la langue amazighe. Pour ce faire, nous nous inspirons des schémas et de l'architecture stratificationnelle ou séquentielle (en série) adoptée par (Sabah, 1988).

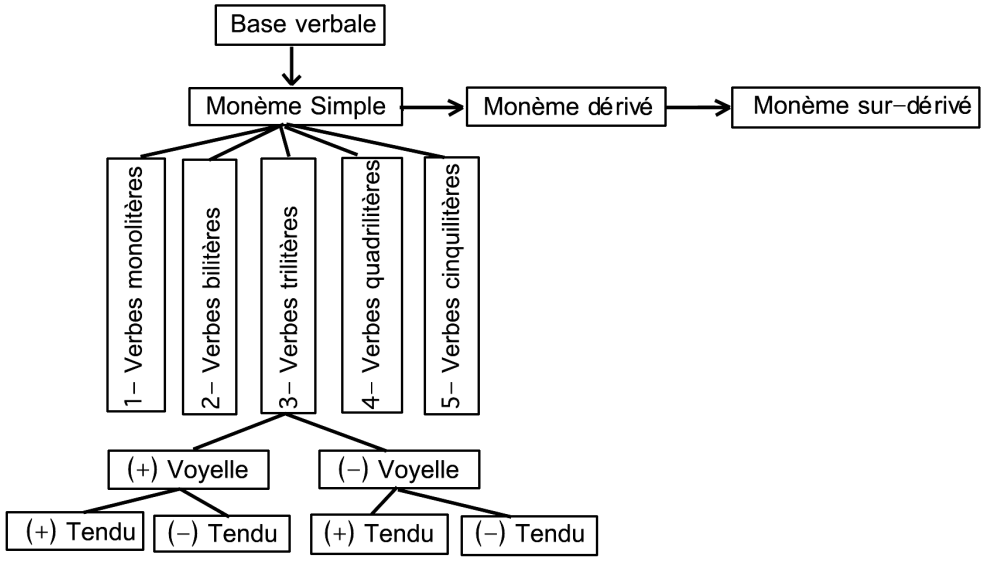


Figure 1: Classification des monèmes verbaux

5.2.2. Schéma basé sur l'opposition thématique

Lorsque nous parlons d'oppositions thématiques (valeurs aspectuelles), nous parlons certainement des différences morphologiques enregistrées aux différents thèmes verbaux. Ces variations morphologiques peuvent être exploitées pour la classification et la taxinomie des verbes. Autrement dit, faire un classement suivant le nombre des thèmes verbaux susceptibles de générer un verbe. Néanmoins, la contrainte qui se pose est la détermination, lors de passage d'un thème verbal à l'autre, de type d'opposition thématique (aoriste vs accompli, accompli vs inaccompli, etc.).

En revanche, il ne semble pas que tous les verbes admettent cette analyse, quoique, sur le plan formel, la plupart des verbes donnent leurs intensifs par gémination d'une consonne ou la préfixation d'un tt-. D'autre part, pour sa part, le prétérit est majoritairement marqué par une alternance vocalique.

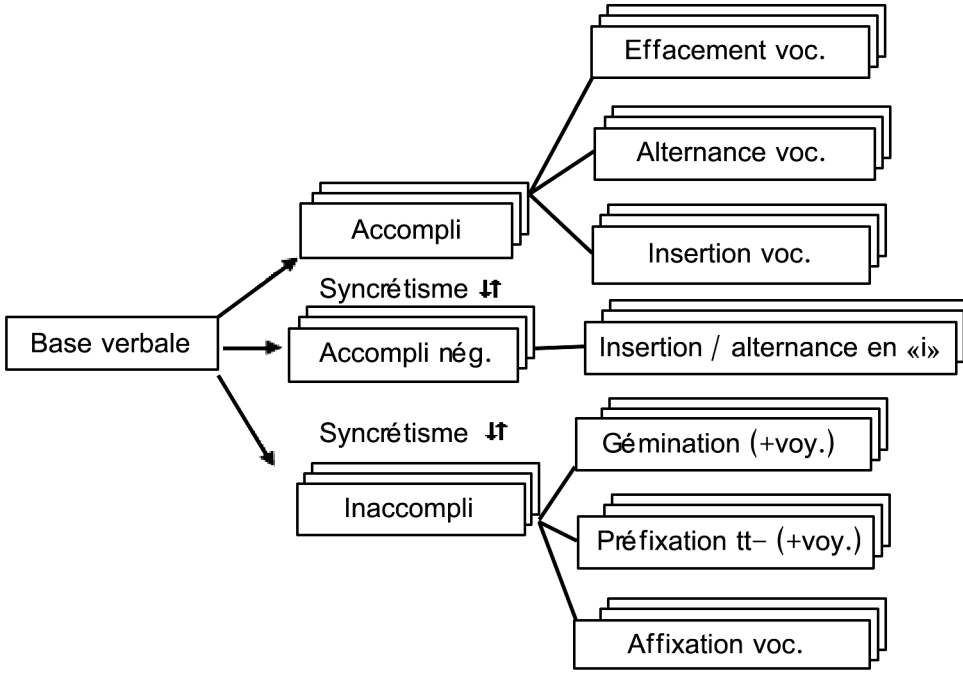


Figure 2: Formation des thèmes verbaux et opposition thématique

Il arrive de distinguer des sous-classes, à l'intérieur de ces classes, en fonction des thèmes verbaux qui expriment un syncrétisme. Néanmoins, certains verbes peuvent exprimer un syncrétisme, selon les parlers, pour certains thèmes verbaux. Par conséquent, un verbe qui présente des différences morphologiques au niveau des thèmes habituels serait classé, selon le nombre de ces oppositions thématiques. Ainsi, un verbe à quatre thèmes serait le verbe qui présente des variations lors de la formation des quatre thèmes. Il est de même pour les verbes à trois, à deux et même à thème unique qui présente un syncrétisme (*awra* « venir » ; *ms* «être» syn. g). Ce syncrétisme nous renseigne évidemment sur le nombre des thèmes d'un verbe donné (*cf.* Boukhris 1986, pour tamazight et Dallet 1953, pour le kabyle).

5.2.3. Schéma basé sur le schème ou la structure des monèmes verbaux

Ce critère est basé principalement sur la constitution, la disposition et l'assemblage des éléments qui forment l'ossature de la forme de base, c'est-à-dire sur l'agencement des phonèmes consonantiques et vocaliques dans chaque structure verbale. Les caractéristiques de chaque structure dépendent du type verbal et du schème qui lui donne sa cohérence.

Ce critère est basé sur les processus investis dans la conjugaison. Il ne se base pas sur la structure verbale étant donné qu'il existe des manières de conjugaison différentes pour une même structure verbale. D'ailleurs, les procédés de la mise à la forme conjuguée sont toujours les mêmes pour tous les parlers. Ces critères ou procédures de classification sont basés sur

les procédés de formation des thèmes verbaux[13] ou sur la façon de conjugaison[14]. Surtout, les procédés investis dans la conjugaison, à savoir la variation consonantique et vocalique. Néanmoins, leur application n'est pas toujours la même.

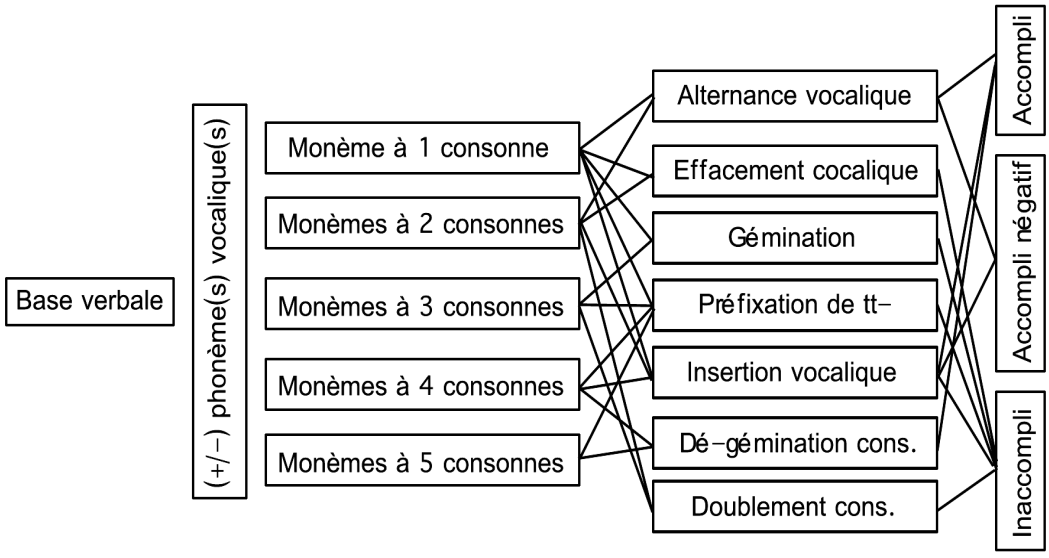


Figure 3 : Taxinomie partielle de verbes

La formalisation des données peut se faire suivant le type de conjugaison ou de procédés de formation. Sans prendre en considération la nature du syncrétisme réalisé entre les thèmes verbaux, le problème majeur reste le comportement morphologique hétéroclite de ces verbes selon les thèmes verbaux[15]. D'ailleurs, nous distinguons des verbes dont les thèmes sont caractérisés par des alternances ou des insertions vocaliques et des verbes qui présentent un effacement de la voyelle dans certains de leurs thèmes. D'autres éléments peuvent intervenir dans la formation des thèmes ; en l'occurrence, la tension consonantique, la perte de la tension, l'alternance ou l' insertion vocalique.

13 Pour Bentolila, le thème verbal serait l'équivalent de la base verbale ou du verbe sans affixe (Bentolila, 1981).

14 Il est mieux de retenir que la mise à la forme de l'aoriste est parfaitement simple et s'obtient par l'affixation des indices de personne à la base verbale dans la presque totalité des parlers. Autrement, ces conjugaisons sont obtenues par l'addition des désinences qui font corps avec le radical verbal. Ils peuvent être préfixés ou suffixés et ils s'ajoutent au verbe quel que soit sa forme et indépendamment de temps.

15 En gros, si le radical de l'aoriste est de la forme /ccc/ ou /ccv/, on gémine sa deuxième consonne, mais dans quelques cas, on lui préfixe /tt/. Si le radical de l'aoriste est de la forme /cc/, on gémine sa première consonne à l'inaccompli, et on insère la voyelle « a ». (*gn/ggan* ; *fl/ffal*) à l'exception de certains verbes qui ne sont pas sujets à l'insertion de « a » ; dans ce cas, on gémine leur deuxième consonne : *ṭṭf/ tteṭṭf*. À côté de ces formes qui préfixent un « *tt-* », il existe une catégorie qui préfixe « *tti-* » et perd la tension de la consonne tendu de thème d'aoriste (*ddu - ttudu*) au centre.

La formation de l'accompli se caractérise par un syncrétisme de forme avec le thème de l'aoriste, ou bien par des alternances vocaliques s'affichant sous forme de variations, changement ou effacement du phonème vocalique. Par contre, la formation de l'inaccompli peut être réduit à trois procédés majeurs. Premièrement, la préfixation du morphème tt- généralement quand le verbe commence par une voyelle ou par une consonne géminée (*asy/ttasy ; ini/ttini; ddu, ttddu*, mais aussi *ttudu*). Ce procédé d'adjonction de la consonne dentale « *tt-* » ou « *t* » à l'initiale concerne d'autres formes verbales. Deuxièmement, l'affixation d'un phonème vocalique. Troisièmement, la gémination d'une consonne radicale constitutive de la base. D'une certaine manière, les deux derniers procédés morphologiques sont combinés dans beaucoup de cas, parfois, les deux premiers procédés, tandis que la gémination n'est presque jamais combinée avec la préfixation de « tt ».

En revanche, la difficulté de la conjugaison porte surtout sur les modifications structurelles affectées par l'alternance vocalique pour garder une bonne sonorité. Encore, certains parlers ont régularisé la règle de l'insertion vocalique en position post-finale et d'autres ont un syncrétisme de forme accompli et inaccompli. Ce critère n'est pas a priori différencié puisque la différenciation des classes ne se base pas sur un regroupement des cas distincts.

6. Conclusion

Les verbes peuvent être classés selon le nombre de phonèmes constituant les verbes en groupes ou catégories. Ce critère ne peut être qu'un raisonnement trompeur parce qu'un classement de ce type n'est pas vraiment pertinent et homogène. Certains verbes peuvent être constants au prétérit et à l'aoriste; d'autres, au prétérit ou à l'aoriste seulement. Vu l'hétérogénéité enregistrée, le classement ne doit pas utiliser la forme nue du verbe. La règle est assez simple en apparence ; mais, elle pose en profondeur le problème de conflit des critères.

La taxinomie des connaissances morphologiques rend problématique la détermination d'une forme de base pour les lexèmes verbaux, d'une part, et d'une autre, des alternances morpho-phonologiques intervenant dans la conjugaison. Nous supposons, pour mieux classer un monème verbal, de le mettre en relation avec ses formes concurrentes. Le repérage et l'extraction automatique des internements ont été à la base de toutes les études de construction automatique.

Il existe deux types de travaux sur la complexité linguistique : les uns s'attachant à décrire les facteurs de complexité et les autres proposant de prédire les zones de complexité observées pendant le traitement. Nonobstant, les méthodes de classification utilisées pour classer des verbes en groupes homogènes sont formellement diverses. Pour ces raisons, nous entrevoyons une utilité de la classification automatique. Malheureusement, les approches taxinomiques ne sont pas exploitables en l'état et demandent une manipulation qui peut devenir très fastidieuse étant donné l'ampleur de la tâche. Les approches sont trop génériques pour appréhender les relations morphologiques impliquées par les thèmes verbaux.

Il faut lancer la recherche dans ce domaine de la taxinomie pour évoluer favorablement vers la résolution des relations morphologiques permettant d'obtenir des classes moins nombreuses et plus homogènes. Les approches de classification peuvent être améliorées dans des cadres applicatifs. Par conséquent, les démarches récentes de la taxinomie des langues naturelles se

basent sur les outils du traitement automatique tels que les *analyseurs morphologiques*, les *concordanciers et les racineurs*, entre autres. D'où la nécessité d'emprunter le même chemin pour le développement informatique de la langue amazighe.

Références

Aspinion R. (1953). *Apprenons le Berbère, Initiation aux dialectes cheleuhs*, Rabat, Moncho.

Basset A. (1929). *La langue berbère. Morphologie. Le verbe. Etude de thèmes*. Paris : Leroux.

Bentolila F. (1981). *Grammaire fonctionnelle d'un parler berbère. Aït Seghrouchen d'Oum Jeniba (Maroc)*. Société d'Etudes Linguistiques et Anthropologiques de France. Paris, pp. 60.

Bouillon P. (1998). *Traitement automatique des langues naturelles champs linguistiques*, Editions Duculot, Paris, Bruxelles.

Boukous A. (1987). *Phonotactique et domaines prosodiques en Berbère (parler tachelhit d'Agadir*. Thèse de Doctorat d'Etat. Vincennes à Saint-Denis: Université de Paris VIII.

Cantineau J. (1950). Racine et schème. *Mêlanges William Marçais*, Paris, Maisonneuve.

Cardey S. et Mandic R. (2005). *La traduction et le Traitement Automatique des Langues*, Centre TESNIERE, Bulag, Revue Annuelle, N° 25.

Constant M. *et al*. (2008). *Description linguistique pour le traitement automatique du français*, Cahiers du Central 5, Presses de l'Université Catholique de Louvain,

Dallet J. M. (1953). *Le Verbe Kabyle: 1. Formes simples*, Fichier de documentation berbère. Algiers: Fort-National.

Dell F. et El Madlaoui M. (1987). *Clitic ordering, morphologie and Phonogy in the verbal complex of Imdlawn Tashlhit berber*, Langue Orientales Anciennes, Philogie et Linguistique (LOAPL), 2 et 3.

Dubois *et al*., (1994). Dictionnaire de linguistique et des sciences du langage, Larousse, Paris, pp. 76-86,

El Gholb L. (2011). *La conjugaison du verbe en amazighe : élément pour une organisation*, Editions Universitéaires Européennes, Sarrebruck, Allemagne.

El Gholb L. (2011). *Modélisation de la conjugaison du verbe en amazighe*: quelles contraintes?Actes des IX Rencontres des Jeunes Chercheurs en Parol, 25-27 mai 2011, Université Stendal, Grenoble.

Galand L. (1977). *Etudes de linguistique berbère*, Leuven-Paris, Peeters (Collection linguistique publiée par la Société de Linguistqiue de Paris, 83, pp. 984.

Galand L. (2002). *Continuité et renouvellements d'un système verbal* : le cas du berbère. Bulletin de la Société de Linguistique de Paris. T. LXXII, Fascule 1, 275, 303.

Galand L. (2010). *Regards sur le Berbère*, Studi Camito-Semitici 8, Centro Studi Camito-Semitici, Milano.

Gleason H. A. (1969). *Introduction à la linguistique*, trad. de F. Dubois-Charlier, Paris : Larousse, chap. 7: "Classement des allomorphes en morphèmes", pp. 65-75.

Iazzi E. (1991). *Morphologie du verbe en Tamazight (Parler des Aït Attab, Haut-Atlas Central)*. Approche prosodique, Thèse de DES. FLSH, Rabat.

Lahrouchi M. (2005). *Regular and Irregular Imperfective conjugaisons in Berber languages*, IGG XXXI, February 24-26, Univeristé de Rome.

Laoust E. (1918). *Etude sur le dialecte Berbère des ntifa,* Paris : Leroux, pp.123

Malmberg B. (1979). *La Phonétique*, Paris: PUF- Coll. Que sais-je n°637.

Martinet A. (1985). Syntaxe générale, Paris, Armand Colin, pp.108

Sabah G. (1988). *L'intelligence artificielle et le langage*, volume2, Paris, Hermès.

Svetla K. *et al*. (2007). *Formaliser les langues avec l'ordinateur*, De INTEX à Nooj, Les Cahiers de la MSH Ledoux, Presses universitaire de Franche-Comté.

Xanthos A. (2008). *Apprentissage automatique de la morphologie*, *le cas des structures racine-schème*, Peter Lang SA, Editions scientifiques internationales, Berne.

Projet GCAM
Vers une Gestion Informatisée du Corpus Amazighe à l'IRCAM

Youssef Ait Ouguengay **Amal El Hamdaoui**
Brahim El Hasnaouy **Abdellah Faddah**
Institut Royal de la Culture Amazighe
{ouguengay, elhamdaoui, elhasnaouy, faddah}@ircam.ma

Résumé

La présente contribution présente un travail qui rentre dans le cadre du projet de l'Institut Royal de la Culture Amazighe visant la collecte et la mise en réseau du premier noyau d'une banque de corpus littéraires amazighes. Quelques années sont passées depuis le lancement de l'action de la collecte prise en charge par le centre CEALPEA de l'IRCAM, une nouvelle phase s'est engagée, celle de l'informatisation de ces ressources riches et extrêmement importants pour les chercheurs dans le domaine de la littérature amazighe. Un système de gestion de contenus (CMS) a été développé dans le cadre de ce projet. Les corpus textuels et audiovisuels collectés sont gérés par le CMS et consultables à travers un moteur de recherche intégré dans ce dernier et destiné à être mis à la disposition des utilisateurs. Le travail d'informatisation effectué a consisté en premier lieu en une étude des caractéristiques principales du corpus collecté. Dans un deuxième lieu, la phase de la conception a visé la mise en place du modèle conceptuel de la base de données du corpus et de la logique retenus pour les règles et les fonctionnalités de gestion et de la consultation. La mise en œuvre du projet a été couronnée par le développement informatique de la première version du système de gestion informatisée du corpus amazighe.

1. Introduction

La collecte et le traitement des corpus de la littérature amazighe ont constitué une des actions phares de l'Institut Royal de la Culture Amazighe (IRCAM) depuis le démarrage de ses travaux. L'IRCAM a accordé une importance considérable à la collecte et l'enregistrement de patrimoine oral amazighe. Compte tenue de la disparition et l'extinction de plusieurs créateurs de ce patrimoine : poètes, conteurs, chanteurs, musiciens et écrivains…, la collecte et la préservation des expressions orales amazighes sont devenues un chantier a part entière qui nécessite un encouragement et un soutien moral et financier, d'où la mobilisation d'un grand nombre de chercheurs et de spécialistes pour recueillir et enregistrer ce patrimoine, notamment à l'IRCAM.

Cette initiation qui entre dans le cadre des efforts consentis pour la préservation du patrimoine amazighe orale vient consolider les nouvelles fonctions que la langue amazighe

est appelée à couvrir tels que le passage à l'écrit, l'introduction dans le système éducatif et particulièrement son inscription dans le nouveau statut officiel du Royaume. La présentation des corpus sous le support technologique fait partie des objectifs et besoins évidents des nouveaux modes de recherche et de communication. Ce qui donnera, sans doute, un nouvel élan à cette action de la collecte : le corpus collecté est un corpus ouvert qui est appelé à devenir encore plus important ce qui fait appel à l'outil informatique pour la facilitation de son traitement et sa préservation selon les normes préconisées dans la matière.

Nous proposons de présenter, dans ce papier, un état des lieux du projet de gestion des corpus amazighes (GCAM) dont a fait l'objet le développement d'un système informatique de gestion des corpus, particulièrement littéraires. Dans la première section, nous exposons les caractéristiques majeures du corpus en question avant de développer les phases de la conception et de développement informatique du projet.

2. La collecte des corpus à l'IRCAM : état des lieux

Depuis le début de la collecte, supervisée par le Centre des Etudes Artistiques des Expressions littéraires et de la Production audiovisuelle (CEAELPA), ce dernier a reçu un nombre considérable de corpus collectés et menés par des chercheurs et des spécialistes contractuels. Les textes et enregistrements audio ainsi collectés, contiennent divers genres littéraires, narratifs et artistiques orales amazighes, tels que : la poésie, la prose, les énigmes, les proverbes, les devinettes et les contes. Ce patrimoine littéraire représente les 3 régions du Maroc : Nord, Centre et Sud.

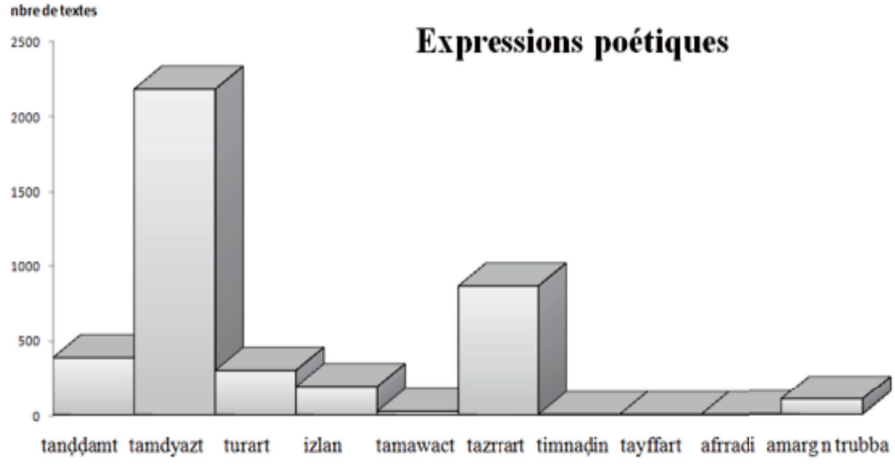


Figure1 : Composition du corpus collecté par genre (poésie)

Le corpus amazighe orale disponible au CEAELPA montre que la poésie et ses sous genres (Figure1) sont les produits les plus collectés, suivie par la prose et ses genres : proverbes,

énigmes et devinettes. Le bilan réalisé par le CEAELPA[1] a démontré la prépondérance des textes poétiques ; Il semble que Tamdyazt ou Amarag ou Aqsid occupe la première place, suivi par Tandamt n uhwacc dite joutes oratoires.

Par ailleurs, les textes en prose représentent la plupart de la collecte. Les proverbes sont la forme majoritaire de la prose (Figure 2), tandis que les énigmes et les contes viennent dans une proportion beaucoup moins importante.

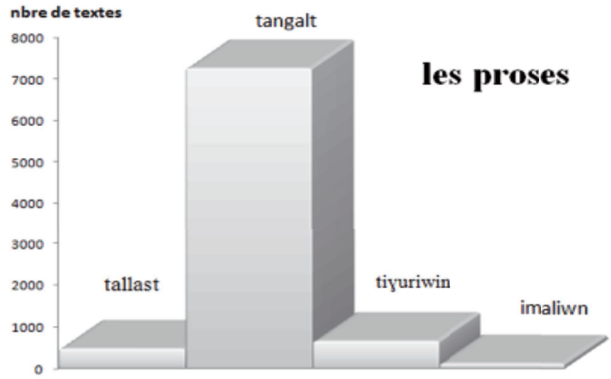


Figure 2 : Composition du corpus collecté par genre (Prose)

En termes de poésie, le nombre total des poètes cités dans la collecte est de 163 dont la majorité est issues de la région de sud (Figure 3), suivi par les poètes du centre, tandis que la poésie du nord appartient à des poètes anonymes.

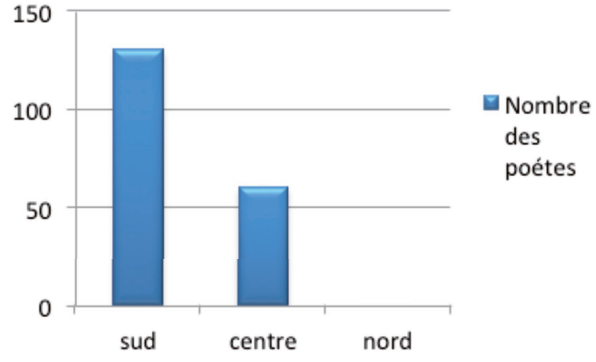


Figure 3 : Nombre de poètes par régions

1 Centre des études littéraires des expressions artistiques et de la production audiovisuelle (CEAELPA) : Bilan de la collecte de patrimoine orale amazighe de 20012002/ au décembre 2007, réalisé par Ahmed El Mounadi, 14 janvier 2008.

Le patrimoine littéraire de la région du nord est menacé plus par la perte et la disparition. Les chiffres avancés montrent que la poésie du sud est la plus représentative dans cette collecte de l'IRCAM. La répartition de ce produit poétique par région montre également une forte présence de la poésie de sud, ceci dit que la collecte littéraire dans cette région est d'une conséquence particulière. Cette présence est due également aux efforts remarquables menés par les chercheurs et les spécialistes en patrimoine amazighe dans ladite région.

2.1. Première exploitation de l'archive

En termes d'exploitation du corpus collecté, plusieurs travaux de publications ont été édités : d'une part une série d'anthologies dont l'anthologie de la poésie, des proverbes et des devinettes (Azdoud 2012-a, Azdoud 2012-b). De même, plusieurs ouvrages qui portent sur la production poétique de certains poètes, tel que (Azeroual, 2012 ; El Mounadi, 2012; Azdoud, 2012-c), ont trouvé dans le corpus une source riche. Le processus d'exploitation de ce corpus continue également pour accomplir l'anthologie des contes amazighs sur laquelle un travail est mené actuellement au CEAELPA.

La demande externe de cette collecte a aussi émergé, notamment de la part des étudiants et des professionnels de la radio et de la télévision. Le projet de numérisation et de gestion de corpus amazighe a pour but, justement, de diffuser ce patrimoine au niveau national et à l'étranger et de permettre aux chercheurs, étudiants et spécialistes de la littérature en générale et de la littérature amazighe en particulier la consultation et l'exploitation de ce patrimoine littéraire.

3. Le projet GCAM

Le corpus littéraire collecté est constitué des documents de natures divers (texte, audio et vidéo). Il nécessite une approche de gestion facilitatrice de son exploitation. La plateforme technologique de gestion du contenu permettrait une gestion unifiée de ces données en visant plusieurs objectifs, notamment :

- Gestion du circuit de l'acquisition du corpus. Incluant toutes les informations concernant le processus de traitement (contractant, rapporteurs, corpus et media, artistes,…).
- Création de corpus et de notices textuelles décrivant chaque document de la base des corpus.
- Mettre à la disponibilité des chercheurs les fonctionnalités de recherche et de navigation dans la base de données.
- Informations et statistiques sur les corpus intégrés.

Dans la démarche de développement adopté, les utilisateurs pilotes ont été impliqués depuis le début de la réalisation. Le développement a été incrémental et itératif de manière que chaque cycle donne lieu à une partie opérationnelle de l'application.

3.1. Mise en œuvre du projet GCAM

Le développement de l'application a été procédé depuis le départ pour pouvoir évoluer en fonction des besoins spécifiques, et intégrer de nouveaux modules et fonctionnalités à la

demande. La mise en œuvre a concerné le développement des fonctionnalités de gestion et la conception de la base de données.

Dans un premier volet, la mise en place de la base de données du projet a été divisée en deux étapes :

3.1.1. Etudes des données

Afin de définir le contenu de la base de données, une étude de la nature des données brutes existantes a été faite à la base du descriptif déjà établis (El Mounadi, 2008). Les documents textuels étaient stockés pour la pluparts sur des fichiers Word avec différents graphies (amazighe, latin et arabe)[2]. Des exceptions ont été notées, qui existent sous format papier et ont besoin préalablement de la numérisation. Les documents multimédias existent sous différents formats audio, vidéo et image. Un dictionnaire de données qui est le fondement de la base de données a ainsi pu être créé à partir de cette étude.

3.1.2. Structuration des données

A partir du dictionnaire, il a été possible de structurer la base de données. Le schéma des données prévoit une description détaillée des différents acteurs et objets de la collecte, à savoir d'une part les informations concernant les auteurs, les contractants et les rapporteurs et d'autre part, les corpus, notamment de type médias (enregistrement audio et vidéo) sont pris en charge, en plus des corpus textuels. La base de données spécifie les éléments de description de chaque objet auteur, etc et explicite les relations entre les différentes tables de métadonnées.

Le modèle conceptuel de la page suivante (Figure 4) résume les caractéristiques relationnelles des données.

3.1.3. Les fonctionnalités développées

Dans le cadre de la création du système, en sus de la base de données, la mise en place de fonctionnalités nécessaires pour l'exploitation et gestion de corpus a été abordée aussi. Ces fonctionnalités sont des leviers techniques qui ont permis de mettre en place des interfaces plus ou moins complexes sur le site Web et offrir une expérience utilisateur encore sous l'expérimentation (ex : formulaires de recherche, demande d'information,...).

L›ensemble des fonctionnalités ci-dessous ont été développées sous formes de modules séparés qui essayent de répondre aux besoins exprimés lors de la conception et la réalisation de l'application Web. Nous citons :

- *Gestion des artistes* : permet la gestion des fiches des acteurs (auteur, artiste, rapporteur, etc.), la liste des corpus collectés et la liste des Médias collectés.
- *Gestion des utilisateurs* : gère la liste des utilisateurs, la fiche de l'utilisateur, les droits d'accès, la langue préférée et le changement du mot de passe.

2 L'unification de la graphie est un souci qui devait être traité avant de l'intégration des corpus dans la base de données. Elle a fait, donc préalablement, l'objet de travaux de transcription.

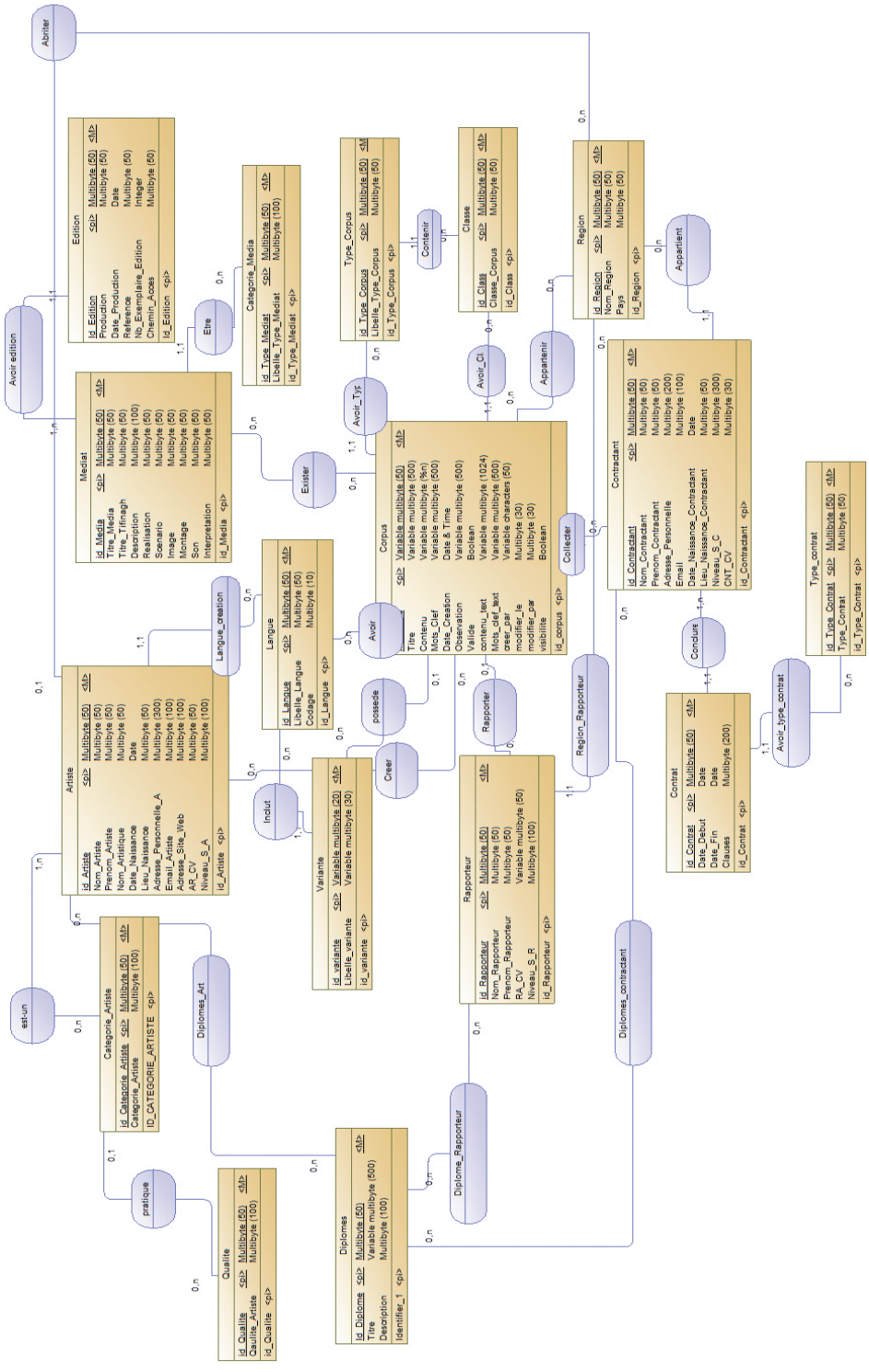


Figure 4 : Modèle conceptuel des données

- *Gestion du corpus* : un module qui permet la saisie, l'édition et la suppression des corpus textuels.
- *Gestion du Média* : Fiche média, ajout des médias, Edition des médias, associer les corpus
- *Paramétrage* : permet la configuration des différents éléments caractérisant le corpus tel que la région, la langue, le type du corpus, le type du média, la classe du corpus, la catégorie artiste, la qualité corpus et le type du corpus.
- *Statistiques* : la version actuelle se limite aux statistiques concernant le nombre de corpus par classe, type, région, contractant ou langue, le nombre d'artiste par région et le nombre de contractant par région.

3.1.4. Mode d'utilisation du GCAM

L'accès aux différents modules de GCAM est profilé selon la fonction à accomplir. Les agents de saisie qui s'occupent de la saisie et l'indexation des corpus, les modérateurs qui valident les actions des agents de saisie et les administrateurs qui ont des droits exclusifs et peuvent paramétrer entièrement le système.

Les formulaires élaborés pour la base de données sur les corpus textuels et médias, ont été conçus pour améliorer la saisie des données. Ils ont été créés dans une optique d'une multi-utilisation de la base, et se veulent donc accessibles pour que les différents acteurs, agents de saisie et modérateurs, amenés à indexer des informations soient le plus rapidement performants.

Figure 5 : Formulaire de recherche

La fonctionnalité de recherche est ouverte à tous les utilisateurs, authentifiés ou non. Elle offre une interface simple et une autre plus riche en matière de métadonnées d'indexation qui permet la recherche dans les corps intégrales des corpus mais aussi dans les données relatives (Figure 5).

4. Conclusion

L'accès aux ressources numériques amazighes est l'une des priorités de la renaissance technologique que connait l'amazighe, notamment les corpus textuels et multimédias qui font partie intégrante des travaux de recherche dans les différents domaines littéraire, historique et anthropologique, etc. La solution développée et exposée dans ce papier répond aux besoins des chercheurs en matière d'accès aux archives corpus. La solution prévoit plusieurs fonctionnalités couvrant la gestion des corpus et des utilisateurs, le paramétrage et la recherche dans la base de données. Les formulaires et interfaces de la consultation des corpus et médias sont dédiés aussi bien aux chercheurs qu'au grand public.

Le projet GCAM est réalisé sous un format modulaire. Il est ouvert aux développements qui répondraient à des nouvelles besoins qui ne surviendront qu'avec l'utilisation de cette version du système et aiderons par ailleurs à améliorer le projet et conduire vers de nouvelles pistes d'évolution.

Références

Azdoud D. (2012). *Anthologie de la poésie amazighe*, IRCAM.

Azdoud D. (2012). ⵜⴰⵡⵉⵙⵉⵜ ⵏ ⵜⴰⵎⴰⵣⵉⵖⵜ, IRCAM.

Azdoud D. (2012). ⴰⵣⵎⴰⵔ ⴰⵏ ⵙⵙⵎⴰⵙⵙⵓⵍ ⵏ ⵜⴰⵎⴰⵣⵉⵖⵜ, IRCAM.

Azeroual F. (2012). ⵜⵉⵏⵎⵎⴰⵔ ⵏⵓⵣⵉⵔ, IRCAM.

El Mounadi A. (2012). ⵜⴰⵡⵓⴰⵜ ⵜⴰⵙⵓⵔⵎⴰⵜ ⵏ ⵜⵎⴰⵢⴰⵙ ⵙⵜⵎⴰⵏⵉ ⵜⵙⵎⵎⵓⵔⵜ, IRCAM.

El Mounadi A. (2008). « تقرير وصفي لحصيلة تدوين التراث الشفوي الأمازيغي منذ تأسيس المعهد الملكي للثقافة الأمازيغية إلى نهاية 2007 ». Rapport interne.

Etude Statistique sur les Erreurs d'Edition dans la Langue Arabe

Hicham Gueddah[1] **Abdellah Yousfi**[2]
[1] Equipe Télécommunications et Systèmes Embarqués, ENSIAS Rabat
Team, SIME Lab ENSIAS, Université Mohammed V, Souissi
h.gueddah@um5s.net.ma
[2] Faculté des sciences juridiques économiques et sociales, Souissi-Rabat
Université Mohammed V Souissi
Yousfi240ma@yahoo.fr

Résumé

Dans le travail présenté dans cet article, nous avons réalisé un test sur les fautes d'orthographe, et ceci pour connaître les différents types d'erreurs d'orthographe commises dans la saisie pour la langue arabe. Nous avons constaté durant ce test que certaines erreurs de permutation, d'ajout et de suppression se répètent de façon fréquente. Ceci nous a poussés à définir des matrices d'erreur d'édition pour la langue arabe.

1. Introduction

La correction automatique des erreurs d'orthographe est l'un des axes les plus importants dans le domaine du traitement automatique des langues naturelles. La recherche dans ce domaine a débuté dans les années 60 (Kukich, 1992).

La correction orthographique consiste à trouver le mot le plus proche des mots d'un lexique d'une langue donnée, en se basant sur la similarité et la distance inter mots.

Pour la correction automatique des erreurs orthographiques, plusieurs travaux ont été réalisés:

- Les premières études ont été consacrées à déterminer les différents types d'erreurs orthographiques élémentaires, appelées opérations d'éditions (Damerau, 1964) notamment: insertion (ajout d'un caractère), suppression (omission d'un caractère), permutation (changement de position entre deux caractères), remplacement (remplacer un caractère par un autre).
- En se basant sur les travaux de Damerau, Levenshtein (Levenshtein, 1966) n'a considéré que trois opérations d'éditions (insertion, suppression, permutation) et il a définit une distance qui compare deux mots en calculant le nombre d'opérations d'édition qui transforme le mot erroné au mot juste. Cette distance est appelée aussi distance de Damerau-Levenshtein.
- Oflazer a proposé une nouvelle approche appelée « Reconnaissance tolérante des erreurs »[1] basée sur l'utilisation d'un dictionnaire représenté sous forme d'automate à

1 Error-Tolerent Finite-State recognition with Applications to Morphological Analysis and Spelling Correction

états finis. Selon cette approche, la correction d'un mot erroné consiste à parcourir un automate à états finis en calculant pour chaque transition une distance appelée ***cut-off edit distance*** (Oflazer, 1996).

- En s'appuyant sur les travaux d'Oflazer, Savary (Savary, 2003) a proposé une variante de cette méthode avec des modifications majeures, du fait qu'elle écarte le calcul du *cut-off edit distance*. Le principe étant d'explorer un dictionnaire-automate en largeur sans dépasser un seuil d'erreurs, et dans le cas de blocage, l'algorithme tolère un retour en arrière (back tracking) pour un parcours en profondeur en se basant sur des suppositions liées aux opérations d'éditions (insertion, inversion, omission, et remplacement).
- Pollock et Zamora (Pollock et Zamora, 1984) ont défini une autre manière pour modéliser les erreurs orthographiques et ceci par le calcul de ce qu'on appelle l'alpha-code (skeletton Key), d'où la nécessité d'avoir deux dictionnaires : un dictionnaire des mots et un pour les alpha-codes. Ainsi pour corriger un mot erroné, on extrait tout d'abord son alpha-code (abréviation des consonnes constituant le mot erroné) et on cherche les alpha-codes les plus proches. Cette méthode est efficace pour le cas des erreurs de permutation.
- Malgré la disponibilité d'un ensemble de méthodes pour la correction orthographique, on constate qu'on ne dispose pas encore de correcticiels[2] robustes capables de traiter de façon pertinente la totalité des erreurs commises dans un texte écrit, voir aussi un mauvais ordonnancement des solutions suggérées lors de la correction. Une analyse critique faite par Souque (Souque, 2006) et Mitton (Mitton, 2009), confirme que les éditeurs de traitement de texte commerciaux tels WinWord et OpenOffice, présentent des limitations dans la suggestion des solutions de quelques types de mots erronés.

Dans cet article, nous avons réalisé une étude statistique sur les erreurs d'édition d'orthographe pour la langue arabe, et ceci dans le but d'introduire les résultats de cette étude dans un système de correction d'erreur orthographique.

2. Etude statistique sur les erreurs d'orthographe

2.1. Influence des paramètres techniques

Pour analyser les différentes erreurs d'orthographe commises par les utilisateurs pour la langue arabe, nous avons effectué un test sur un ensemble d'opérateurs expérimentés en saisie de documents en arabe. Notre test est constitué d'un ensemble de documents écrits en langue arabe avec des polices, tailles et interlignes différents, et ceci pour connaître d'une part l'influence de ces paramètres sur les erreurs d'orthographes, et d'autre part les erreurs d'édition qui se répètent.

2 Le terme correcticiel servira à désigner l'ensemble des outils logiciels qui aident à la correction d'un texte

	Police	Taille de Police	Interlignage
Document A	SimplifiedArbic	12	-
Document B	ArbicTrans	12	1.5
Document C	ArbicTrans	12	-
Document D	ArbicTrans	14	-
Document E	ArbicTrans	14	1.5

Tableau 1 : Caractéristiques techniques des documents

	A	B	C	D	E
Nombre d'erreurs	283	287	308	232	320

Tableau 2 : Statistiques des erreurs orthographiques

D'après les statistiques du tableau 2, nous pouvons dire qu'il y a une relation étroite entre les caractéristiques techniques d'un document Word'et les erreurs orthographiques commises.

Dans ce sens, on peut conclure :

i. Tant que le document est bien saisi (police claire, taille et interlignage adéquats) l'opérateur ne prend'pas assez du temps pour lire attentivement les mots et cela peut entrainer à commettre des erreurs de frappe.

ii. Tant que le document n est pas bien écrit (police, taille et interlignage médiocres) l'opérateur prend'assez de temps pour lire attentivement les mots avant de les saisir ce qui réduit le nombre des erreurs commises mais dans un temps considérable.

2.2. Statistique sur les opérations d'éditions

Les erreurs orthographiques commises dans notre test sont de trois types : ajout, suppression et permutation, appelées aussi opérations d'éditions (Levenshtein, 1966).

a. ***Erreur d'ajout*** : pour ce type d'erreur, l'opérateur de saisie insère incorrectement un ou plusieurs caractères au mot correct (مدرسشة ajout du caractère ش à مدرسة).

b. ***Erreur de suppression*** : dans cette catégorie d'erreur, l'opérateur omet incorrectement un ou plusieurs caractères du mot correct (مدسة suppression du caractère ر).

c. ***Erreur de permutation*** : cela signifie que l'operateur remplace ou bien permute un ou plusieurs caractères dans un mot correct (مدرشة permutation entre ش et س du mot مدرسة).

Opération d'édition	Suppression	Ajout	Permutation
Taux	20,77%	14,23%	65%

Tableau 3 : Pourcentage par opération d'édition

D'après ce tableau, on constate clairement que les erreurs de permutation sont les plus fréquemment commises suivies des erreurs de suppression et des erreurs d'ajout.

3. Analyse des erreurs

3.1. Analyse des erreurs de permutation

Lors de la comparaison des documents d'origine avec ceux saisis par les opérateurs, on a remarqué qu'il y a des caractères arabes qui se fréquentent souvent en erreur de permutation par rapport aux autres. On a comptabilisé un total de 923 erreurs de permutations commises par les opérateurs de saisie.

Le tableau suivant résume les principaux caractères qui subissent en erreur de permutation.

Caractères	ا	أ	ي	ل	ر	إ	ب	ن	ج	ت	ق	م	و	ح	س	ى
Nombre de permutation	151	138	80	41	40	39	36	28	26	23	23	22	22	20	20	20

Tableau 4 : Les principaux caractères qui subissent en erreur de permutation

D'après ces statistiques, on constate qu'il y a des caractères qui sont permutés fréquemment par rapport aux autres et ceci peut être expliqué selon deux interprétations :

- La première interprétation est que les opérateurs se trouvent en face du problème de proximité entre les touches du clavier, et au lieu de cibler le caractère correct l'opérateur frappe la touche adjacente à ce dernier. Par exemple, le caractère ب a été permuté souvent avec les caractères ل , ي , ر et ق qui se trouvent juste à côté du caractère ب (voir Figure 1).

Figure 1 : Clavier alphabétique arabe

- La deuxième interprétation, c'est que les opérateurs se trouvent dans une situation de ressemblance entre les caractères arabe, et par conséquent au lieu de cibler le caractère en question, l'opérateur saisi le caractère qui ressemble à ce dernier. A titre d'exemple, ي —> ى, ر —> ز , ج —> ح , س —> ش , ق —> ف.

3.2. Analyse des erreurs de suppression

Durant cette étude statistique, on a constaté que les opérateurs de saisie ont commis en plus des erreurs de permutation, des erreurs de suppression des caractères avec une somme de 295 erreurs.

Dans le tableau suivant on a essayé de recenser les différents caractères qui ont présenté assez de suppression par rapport aux autres.

Caractères	ا	ت	ر	ك	ل	م	ن	و	ي
Nombre de suppression	53	24	28	10	22	24	15	28	37

Tableau 5 : Les caractères présentant plus de suppression

3.3. Analyse des erreurs d'ajout

La dernière catégorie des erreurs d'orthographe commises par les opérateurs de saisie est l'erreur d'ajout avec un total de 202 erreurs. Le tableau ci-dessus résume les différents caractères sujets d'erreur d'ajout.

Caractères	ت	غ	ر	ن	ل	ك	ا	ي
Nombre d'ajout	12	10	19	19	32	8	23	26

Tableau 6 : Les principaux caractères présentant un sujet d'erreur d'ajout

4. Conclusion et Perspectives

Dans ce travail, nous avons essayé de mettre en évidence certains caractères de la langue arabe subissant souvent un nombre assez important d'erreurs d'édition (ajout, suppression, permutation). Ceci nous a aidés à définir ce que nous avons appelé des matrices de fréquence d'erreur d'édition et à utiliser ces dernières dans la distance de Levenshtein (Levenshtein, 1966) pour mieux corriger les erreurs d'orthographe (Gueddah *et al*., 2012).

Référence

K. Kukich (1992). *Techniques for Automatically Correcting Words in Text* . ACM Computing Surveys, vol. 24, No.4, pp. 377-439, December 1992.

F. J. Damerau (1964). A technique for computer detection and correction of spelling errors. Communications of the Association for Computing Machinery.

V. Levenshtein (1966). Binary codes capable of correcting deletions, insertions and reversals. SOL Phys Dokl, pp. 707-710.

K. Oflazer (1996). *Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction*. Computational Linguistics archive, vol. 22 Issue 1, pp. 73-89, March 1996.

A. Savary (2000). *Recensement et description des mots composés méthodes et applications*. version 1 - 24 Sep 2011, pp. 149-158.

J. J. Pollock, A. Zamora (1984). Automatic Spelling Correction in Scientific and Scholarly Text, Communications of the ACM, 27(4), pp. 358-368.

M. Ndiaye, A. V. Faltin (2004). *Correcteur Orthographique Adapté à Apprentissage du Français*. Revue Bulag n°29, pp. 117-134.

A. Souque (2006). Approche critique des produits IdL: analyse comparative des correcteurs orthographiques de Word 2000 et OpenOffice 2. Master 1 Industries de la Langue, Université Stendhal-Grenoble 3.

R. Mitton (2009). Ordering the suggestions of a spellchecker without using context. Natural Language Engineering 15 (2), pp. 173-192.

H. Gueddah, A. Yousfi, M. Belkasmi (2012). Introduction of the weight edition errors in the Levenshtein distance. *International Journal of Advanced Research in Artificial Intelligence,* vol. 1 Issue 5. Aout-2012.

Analyseur Morphologique des Mots Arabe en Utilisant le Dérivé et Schème de Surface

Said Iazzi[1] **Abdellah Yousfi**[2] **Mostafa Bellafkih**[3]
iazzisaid@yahoo.fr
[1] Laboratoire GSCM-LRIT, FS, Agdal-Rabat, Maroc.
[2] Equipe ERADIASS, FSJES, Souissi-Rabat, Maroc.
[3] INPT, Rabat, Maroc.

Résumé

Cet article présente un système d'analyse morphologique pour la langue arabe. Ce système est fondé sur les schèmes de surface des mots arabes.

Notre travail dans cet article, vise à traiter les noms dérivés arabes, il est basé principalement sur la construction de la base des données des schèmes de surface de ces derniers, ensuite on a adopté un travail antérieur de (Yousfi, 2010) pour l'analyse des verbes arabes pour traiter les noms dérivés arabes.

Notre approche à été testée sur un corpus de 2400 mots arabes (400 verbes et 2000 noms dérivés), les résultats obtenus sont très intéressants et montrent l'utilité et l'importance de cette approche.

1. Introduction

L'analyse morphologique arabe et l'un des outils qui permettent de résoudre la majorité des problèmes de la langue arabe, elle a été largement utilisée dans plusieurs domaines du traitement automatique des langues naturelles (TALN) tels que la recherche documentaire, les dictionnaires électroniques, les systèmes de marquage, etc.

Plusieurs travaux ont été réalisés dans le but d'élaborer des analyseurs morphologiques de la langue arabe et qui peuvent être regroupés en trois approches (Darwish, 2002 ; Yousfi, 2010):

- ***L'approche symbolique***: Cette approche est basée sur la segmentation du mot en préfixes, infixes et suffixes dans le but d'extraire la racine du mot arabe. Plusieurs analyseurs morphologiques ont été élaborés et qui s'appuient sur cette approche (Darwish, 2002 ; Buckwalter, 2002 ; Hegazi et ElSharkawi, 1986 ; Koskenniemi, 1983; Beesly, 1998 ; El-Sadany et Hashish, 1989 ; Khoja et Garside, 1999 ; Soudi, 2002). Parmi les analyseurs les plus connus pour cette approche est celui de Buckwalter. Ce dernier consiste à déterminer toutes les segmentations possibles du mot, puis à chercher les résultats dans les listes des radicaux, des suffixes et des préfixes, et vérifie ensuite si les morphologies de chacun des éléments sont compatibles entre elles en examinant trois tables de correspondances : préfixe-radical, préfixe-suffixe, radical-suffixe.

- ***L'approche statistique***: Cette approche calcule les possibilités et les probabilités qu'un préfixe, suffixe et un radical peuvent apparaître ensemble dans une base de données des mots (Goldsmith et John, 2001).
- ***L'approche hybride***: Cette approche combine entre les deux approches précédentes (Darwish, 2002).
- Malgré les avantages de ces approches, on remarque qu'il y a toujours des inconvénients pour ces dernières, on cite par exemple:
  - Le dictionnaire des mots est très grand, et il est très difficile de construire un dictionnaire contenant tous les mots arabes. Ces dictionnaires des mots contiennent une sorte de répétition des noms ayant les mêmes règles morphologiques 'شَاءَ ــ مَشِيءٌ' ، 'جَاءَ ــ مَجِيءٌ'.
  - Ces approches utilisent plusieurs règles au moment de l'analyse morphologique. Pour remédier à ces problèmes, nous avons développé un analyseur morphologique indépendant du dictionnaire des mots et n'utilisant pas les règles au moment de l'analyse morphologique. Notre système utilise uniquement les schèmes de surface du mot à analyser.

2. Construction de la base des schèmes de surface des noms dérivés

2.1. Noms dérivés arabe

Les noms dérivés se sont les noms qui peuvent être dérivés à partir d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. Parmi les noms dérivés (Voir Tableau 1), on cite (Mesfar, 2008):

- *Le Participe Actif* 'اسم الفاعل ' : est un nom associé à tout verbe d'action, et qui désigne l'agent du verbe c'est-à-dire celui qui fait l'action. Par exemple, le verbe 'ضَرَبَ' a pour nom actif ' ضَارِبٌ '.
- *Le Participe Passif* 'اسم المفعول ' : est un nom associé à tout verbe d'action transitif. Il désigne le patient qui subit l'action ou le résultat de cette action. Par exemple le verbe 'سمع' a pour participe passif 'مسموع'.
- *Nom Verbal* 'المصدر ' : est un nom abstrait formé sur la même racine que le verbe auquel il est associé et exprime le même contenu sémantique que le verbe. Un verbe peut avoir plus qu'un nom verbal. Par exemple, le verbe 'وَدَّ' admet quatre noms verbaux différents ' وُدًّا ــ وِدَادا ــ وِدادةً ــ مَوَدَّةً '.
- *La qualité similaire* 'الصفة المشبهة' : les noms de la qualité similaire indiquent la présence absolue de la qualité de celui qui a fait l'action, comme 'رؤوف'.
- *Le Nom comparatif* 'اسم التفضيل' : il indique la qualité commune de deux noms dont l'un exprime un degré supérieur, comme 'أخوَفُ'.

- *Les noms de lieux et de temps* 'اسم الزمان والمكان' : ils indiquent l'endroit ou le temps de l'action, comme 'مَأْخَذٌ' ، 'ملعب' .
- Le nom d'instrument 'اسم الآلة' : il indique le moyen par lequel l'action a été réalisée, comme 'ملعقة'.

Dans cet article, en plus de ces noms, on a traité aussi les noms dérivés suivants : nom d'état "اسم الهيئة " ,"المصدر الميمي", masdar sinaai "المصدر الصناعي", hyperbole "صيغة المبالغة" et "اسم المرة".

Racine	Genre et nombre du nom	Type du nom	Dérivation de la racine
ضَرَبَ	مثنى–مذكر	اسم–الفاعل	ضَارِبَانِ
قَالَ	جمع–مذكر	اسم–الفاعل	قَائِلُونَ
وَلِيَ	مثنى–مذكر	اسم–المفعول	مَولِيَّانِ
ضَرَبَ	جمع–مؤنث	اسم–المفعول	مَضْرُوبَاتٌ
أَمُوَ	مفرد–مذكر	الصفة المشبهة	أَمِيّ
وَقَى	مفرد–مذكر	التصغير	وُوَيْقٍ
وَقَى	مفرد– مؤنث	المصدر	وَاقِيَةً
وَجَأَ	مفرد–مذكر	اسم الزمان و المكان	مَوْجأ
وَقَى	مفرد–مؤنث	المرة	وَقْيَةً
رَعَى	مفرد–مذكر	صيغة–المبالغة	رَعَّايٌ
رَمَى	مفرد–مذكر	اسم–الألة	مرمى
رَعَى	مفرد–مذكر	المصدر–الميمي	مَرْعىً
وَقَى	مفرد– مؤنث	المصدر–الصناعي	الوقائيَّةُ
خَافَ	مفرد–مذكر	اسم–التفضيل	أخوَفُ
رَمَى	مفرد–مؤنث	الهيئة	رِمْيَةً

Tableau 1: Un exemple des noms dérivés en fonctions de leurs racines, genre et nombre

2.2. Schème de surface

Le schème d'un mot permet de détecter les lettres constituant sa racine. Le schème de 'مُكْرِمُونَ' est 'مُفْعِلُونَ', les lettres "ف، ع،ل " remplacent les lettres de la racine de 'مُكْرِمُونَ', et le schème de " صَارَعَ " est " فَاعَلَ" (Youssef, 1999 ; Bahrak, 869-930 ; Hanafi, 1914 ; Zanjani, 1343).

Ce type de schème ne peut pas présenter les variations morphologiques du mot (par exemple le nom قَائِلٌ du verbe قَالَ), c'est pourquoi nous avons proposé un schème adapté appelé schème de surface (Yousfi, 2010).

La méthode de construction de ce nouveau schème est la suivante :
Si on suppose que le mot dont on cherche son schème est :

$w = l_1 l_2 ... l_n$ (l_i Caractère du mot w) et R sa racine.

Le schème de surface de w est $p = f_1 f_2 .... f_n$ avec :

$$\begin{cases} f_i \text{ est l'une des trois lettres "ف، ع، ل " si } l_i \in R \\ f_i = l_i \text{ si } l_i \text{ n'est pas dans R.} \end{cases}$$

Et le schème de surface de la racine $R = g_1 g_2 ... g_k$ (g_i est un caractère) est $P = f'_1 f'_2 .... f'_k$ avec :

$$\begin{cases} f'_i = \text{ l'une des trios lettre "ف، ع ، ل " siest une lettre constante au moment de la conjugaison de R.} \\ f'_i = g_i \text{ i sinon.} \end{cases}$$

Exemple :

La conjugaison du mot 'رَعَى' au participe actif à la 1ére personne du singulier, est 'رَاعٍ', alors le schème de surface de la racine 'رَعَى' est 'فَعَى' et 'فَاعٍ' est le schème de surface de 'رَاعٍ'.

Le schème de surface de 'آجِرٌ' est 'آفِعٌ' et de 'آجِرَاتٌ' est 'آفِعَاتٌ'.

Pour la construction de la base des schèmes de surface des noms dérivés arabes, nous avons traité 127 racines qui représentent presque toutes les classes possibles pour générer les noms dérivés arabes (Youssef, 1999).

Des linguistes ont généré tous les noms dérivés arabes à partir de ces 127 racines, ensuite ils les ont conjugué aux différentes personnes (masculin singulier, masculin duel, masculin pluriel, féminin singulier, féminin duel, féminin pluriel), et à partir des ces noms, ils ont dégagé les schèmes de surface de chaque nom dérivé.

A la fin, nous avons obtenu plus de 6216 schèmes de surface qui représentent presque tous les noms dérivés arabes (Voir tableau 2).

Racine	Genre et nombre du nom	Type du nom	Schème de surface du nom dérivé	Nom Dérivé
اِحْتَرَمَ	جمع–مذكر	اسم–الفاعل	مفتَعلونَ	مُحْتَرِمُونَ
رَعَى	مفرد– مذكر	اسم–الفاعل	فَاعٍ	رَاعٍ
خَافَ	مفرد– مذكر	اسم–الفاعل	فَائِع	خَائِف
قَاوَلَ	مفرد– مذكر	اسم–الفاعل	مُفاوِع	مُقَاوِل
وَلِيَ	مفرد–مذكر	اسم– الفاعل	وَافٍ	وَالٍ
وَلِيَ	مفرد– مذكر	اسم–المفعول	مَوْفِيّ	مَوْلِيّ
اِسْتَعَادَ	مثنى–مؤنث	اسم–المفعول	مُسْتفاعَتَانِ	مُسْتَعَادَتَانِ
أَخَذَ	مفرد–مذكر	المصدر	مأفعاً	مأَخْذًا
حَارَ	مفرد–مذكر	اسم–المفعول	مَفِيع	مَحِير
حَارَ	مفرد–مذكر	الصفة المشبهة	فَائِع	حائِر

Tableau 2: Un exemple des schèmes de surface en fonction de leurs racines, genre et nombre

3. L'approche utilisée dans notre analyseur morphologique

Dans l'approche déjà utilisé par (Yousfi, 2010), on a remarqué que pour la construction de la base des schèmes de surface des verbes, elle ajoute une étape de liaison de tous les suffixes et les préfixes possibles avec les schèmes de surfaces des verbes conjugués, ceci rend la taille de la base des données des schèmes assez grande.

Dans notre cas, on a supprimé cette étape et on a intégré dans le système une phase de segmentation du mot en suffixe, et en préfixe avant de trouver le schème de surface de ce mot.

Exemple :

Le mot 'فواقيانهم' après l'extraction du préfixe 'ف' et du suffixe 'هم' on trouve 'واقيان', donc le schème de surface est 'واعيان'.

On cherche les schèmes du mot dans l'ensemble des schèmes du surface ayant la même longueur.

De même pour ce travail, nous avons pu formuler la fonction qui mesure la similarité entre le mot à analyser et les schèmes de surface. Cette fonction a été formulée comme suit :

$$f(m;w)=\sum_{i=1}^{N} 1_{[m_i;w_i]} \qquad \text{avec :} \qquad 1_{[m_i;w_i]} = \begin{cases} 1 & si\ m_i = w_i\ (m_i = ف, = ع, = ل) \\ f(m,w)=0 & sinon\ \ et\ on\ sort\ de\ l'algorithme \end{cases}$$

$m_i : i^{ème}$ Caractère du schème m et $w_i : i^{ème}$ Caractère du mot w.

La fonction *f* dégage un ensemble de solutions de schèmes de surface, qu'on note par S :

$$S = \left\{ \text{m} \in \text{P}_{\text{L(w)}} \quad / \quad f(m,w) > 0 \right\}$$

$\text{P}_{\text{L}(w)}$: l'ensemble de tous les schèmes de surface de longueur L(*w*).

L(*w*) : la longueur du mot *w*.

Exemple :

f ('واقيان' ; 'وافيان') = 6

f ('واقيان' ; 'فاعيان') = 6

f ('واقيان' ; 'متفاعى') = 0

Ensuite, pour chaque schème de surface m_k du mot *w* on cherche ces racines R_{k_r}. Pour trouver les racines du mot à analyser *w*, on cherche dans un premier temps les positions des caractères "ف", "ع", "ل" dans les schèmes de surface du mot *w* et on dégage les caractères associés à ces positions dans le mot *w*. Ces caractères sont remplacés ensuite dans les schèmes de surface de la racine dans leurs positions.

Par exemple, pour le mot قائلون on trouve les schèmes de surface :

فاعلون avec le schème de surface فعل pour sa racine.

فائلون avec le schème de surface فال pour sa racine.

Après l'application de notre méthode on trouve les deux solutions suivantes :

قائلون ⟵ فاعلونَ ــ فعل ⟵ قَتَلَ

قائِلون ⟵ قائلون ــ فَالَ ⟵ قال

Comme la racine قَتَلَ n'existe pas dans la langue arabe, on garde donc seulement la deuxième solution قَالَ. (Voir tableau 3).

Suffixe	Prefixe	Genre et nombre du nom	Type du nom	Schème mot(w)	Schème racine	Racine	Mot(w)	Mot
هم	ك	مثنى–مذكر	اسم–المفعول	مأجوعان	أجع	أجر	مأجوران	كمأجورانهم
هم	ك	مثنى–مذكر	اسم–المفعول	مجعوران	جعر	أجر	مأجوران	كمأجورانهم
هن	فال	جمع–مذكر	اسم–المفعول	مأجوعون	أجع	أجر	مأجورون	فالمأجورونهن
هن	فال	جمع–مذكر	اسم–المفعول	مجعورون	جعر	أجر	مأجورون	فالمأجورونهن

Tableau 3 : Exemple des résultats de l'analyse morphologique des mots

4. La mise en œuvre

Pour tester notre approche, nous avons d'abord construit tous les schèmes de surface des noms dérivés arabe, cette étape a été réalisée par des linguistes, et ils ont utilisé un ensemble de références arabes (Mustapha, 1999 ; Bahrak, 930 هـ ; Hanafi *et al*., 1914 ; Zanjani, 1343).

Pour la mise en œuvre de notre approche, nous avons développé un programme en java constitué des parties suivantes (voir la figure 1).

- Partie 1 : segmenter le mot en suffixes, préfixes et radical.
- Partie 2 : chercher les schèmes de surface des solutions données par la partie 1.
- partie 3 : chercher les racines à partir de tous les schèmes de surfaces retenus par la partie 2.
- partie 4 : vérifier la validité de ces racines en cherchant s'ils existent dans la base des racines ou non.

Cette approche a été évaluée sur une liste de 2400 mots (400 verbes et 2000 noms dérivés). Ces mots sont différents de ceux utilisés dans la phase de la construction des schèmes de surfaces.

Le taux d'erreur global trouvé est de 3.9%, la majorité de ces erreurs provient principalement de l'insuffisance de la base des données des schèmes de surface, il y a des noms dérivés dont leurs schèmes de surface n'existent pas dans notre base des schèmes. Pour le reste des erreurs, il provient de la phase de génération ou de la construction de ces schèmes de surface.

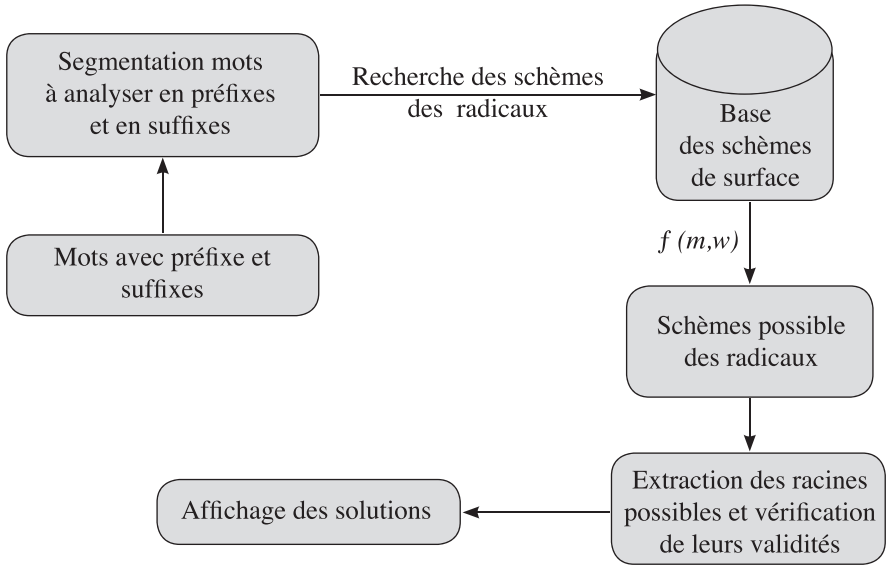


Figure 1 : Les étapes de notre analyseur morphologique des noms dérivés arabe

5. Conclusion

Notre contribution a été d'abord le traitement des noms dérivés arabe qui n'ont pas été traités dans le cas de (Yousfi, 2010), ensuite on a formulé la fonction qui mesure la similarité entre les mots et les schèmes de surfaces. De même on a pu réduire la taille de la base des schèmes en éliminant la phase d'ajout des préfixes et des suffixes aux schèmes de surfaces.

Références

Al-Kharashi I., Evens M. (1994). *Comparing words*, stems, and roots as index terms in an Arabic information retrieval system.

Al Fedaghi S. S., Al-Anzi F. S. (1989). A new application to generate Arabic Root-Pattern Forms, *Procedings of the 11th National Computer Conference and Exhibition, March, Dahran, Saudia Arabia*, pp. 391-400.

Bahrak. (930 هـ) بحرق، جمال الدين محمد بن عمر بن مبارك الحميري الحضرمي , فتح الأقفال وحل الإشكال بشرح لامية الافعال.

Beesly K. R. (1998). Arabic Morphology Using Only Finate-State Operations, *Proceedings of the Workshop on Computational Approaches to Semetic languages. Montreal, Quebec*, pp. 50-57.

Buckwalter.T (2002). Buckwalter Arabic Morphological Analyzer. Version 1.0. *Linguistic Data Consrtium*, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.

Darwish. K. (2002). *Building a shallow Arabic morphologi-cal analyser in one day*. in Proceedings of the ACLWorkshop on Computational Approaches to Semitic Languages, Philadelphia, PA.

El-Sadany T. A., Hashish M. A. (1989). An Arabic Morphological System. *IBM Systems Journal*. vol.28, No.4, pp. 600-612.

Goldsmith, John A. (2001). *Unsupervised learning of the morphology of a natural language*. Computational Linguistics, 27(2), pp. 153-198.

Hanafi N. *et al*. (1914). حنفي بك ناصف–محمد بك دياب–مصطفي طموم–محمود افندي عمر–سلطان بك محمد، *قواعد اللغه العربيه لتلاميذ المدارس الثانويه*، طبعه مصر سنه 1914 اديان. علوم الدين

Hegazi N., ElSharkawi. A. (1986). Natural Arabic Language Processing, *Proceedings of the 9th National Computer Conference and Exhibition, Riyadh, Saudi Arabia*, pp. 1-17.

Khoja S., Garside R. (1999). Stemming Arabic text. *Computer Science Departement, Lancaster University*, Lancaster, UK.

Koskenniemi K. (1983). Two Level Morpology: A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dep. of General Linguistics, University of Helsinki, Helsinki.

Mesfar S. (2008). *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*. Thèse de doctorat, Université de France-comte.

Mustapha G. (1999). *جامع الدروس العربية* –الشيخ مصطفى الغلاييني، المكتبة العصرية

Otakar S. (2007). *Functional Arabic Morphology Formal System and Implementation*. Thèse de doctorat, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague.

Saliba B., Al-Dannan A. (1989). *Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis*. Proceedings of the First Kuwait Computer Conference, Kuwait, March, pp. 3-5.

Sam A., Youssef D. (1999). سام عمار، يوسف ديشي، *مجموعة العربية الأفعال بيشرال*. هاتييه باريس

Soudi A. (2002). A Computational Lexeme-Based, *Treatment of Arabic Morphology*. Doctorat d'état, Mohamed V University.

Yousfi A. (2010). *The morphological analysis of Arabic verbs by using the surface patterns*. *IJCSI International Journal of Computer Science Issues*, vol. 7, Issue 3, No 11, May 2010.

Yousfi A., Sabri S., Bouyakhf E. (2006). *Système d'analyse morphologique des noms Arabes*. JETALA 2006 (Journées d'Etudes sur le Traitement Automatique de la Langue Arabe), Rabat 5-7 juin, 2006.

Yousfi A., Sabri S., Bouyakhf E. (2006). *Système d'analyse morphologique des noms Arabes*. MCSEAI'06 December 07-09, 2006, Agadir, Morocco.

Zanjani . (1343 هـ) الزنجاني . *متن البناء و متن التصريف العربي*، القاهرة مطبعة مصطفى البابي الحلبي

Automatic Tamazight Spelling Correction Using Noisy Channel Model and Bigram Language Model

Said Gounane[1] **Mohamed Fakir**[1] **Belaid Bouikhalen**[2]
[1]Dep. Of Computer Sciences – FST Beni Mellal, Morocco
Gounane.said@gmail.com fakfad@yahoo.fr
[2]Dep. Of Computer Sciences – Multidisciplinary Faculty, Beni Mellal, Morocco
bouikhalen@yahoo.fr

Abstract

In this work, we present an application of noisy channel model and bigram language model to Tamazight spelling auto-correction. Texts are written using Tifinagh character. Tmazight language is modelled by the bigram language model using a small corpus extracted from the IRCAM website. The noisy channel algorithm predicts candidates from the corpus for the misspelled word, then bigram language model assign the appropriate correct word given the tow words surrounding the error in the sentence.

1. Introduction

All modern word processors are using spelling errors detection and correction's algorithms. These algorithms are also used in optical character recognition (OCR), on-line handwriting recognition, to improve their recognition rate, based on a language model and information extracted from text images.

A spelling error could be a non-word error or a real-word error. The first one is detected easily by looking for that word in a dictionary, if it doesn't appear; it's a spelling error of an existing word (real word). From that misspelled word one can generate many hypotheses by a single transformation (deletion, insertion, transposition or replacement) then we chose the most probable word to be mistyped as the misspelled word (noisy channel algorithm). The second type of errors needs the surrounding words to make a decision if the word isn't misspelled as another real word. In that case we use the N-gram language model; a special case is the bigram language model where we use only tow surrounding words of the error.

2. Noisy channel model

Noisy channel algorithm was first proposed by (Kernighan *et al*., 1990). This algorithm is made up of two stages:

1. Proposing candidates correction (c) for the observed error (o) ;
2. Scoring these candidates.

2.1. Minimum edit distance

The minimum edit distance between two words is the minimum number of editions (insertion, deletion, substitution) needed to transform one into the other (Jurafsky et Martin, 2000). For each of these editing operations one can assign a cost. The particular case used in this work is the Levenshtein distance where each of these operations has a cost of 1.

For example the Levenshtein distance between ⵓⵍⵓⵏ and ⵓⵔⵓⵏ is 2 (substitution: one deletion and one insertion), between ⵜⵙⵙⵉⵡ and ⵜⵣⵓⵓⵢ is 7 (three substitution and one deletion).

2.1. Proposing candidates

Since most spelling errors are of minimum edit distance one or two (Jurafsky et Martin, 2000), we suppose that the correct word will differ from misspelled word by single (insertion, deletion, substitution, transposition). Any word in the corpus resulting from a single transformation of the misspelled word is in the candidate set *C*. Word ⵓⵣⵓⵏ could be a misspelled as ⵓⵍⵓⵏ, ⵓⵔⵓⵏ or ⵣⵍⵓⵏ. Scoring system is the next stage to select just one or a list of the most appropriate words.

2.2. Scoring candidates

The score of a candidate word (c) is the probability *P(c/o)* that one can compute using the equation:

$$P(c/o) = \frac{P(o/c)P(c)}{P(o)} \qquad (1)$$

The most likely correction is:

$$\hat{c} = \underset{c \in C}{Arg\max} \frac{P(o/c)P(c)}{P(o)} \qquad (2)$$

Since *P(o)* doesn't depend on c, equation (2) become:

$$\hat{c} = \underset{c \in C}{Arg\max} \left(P(o/c)P(c) \right) \qquad (3)$$

The term *P(c)* is the probability to get the word c from the corpus, this term is obtained just by counting its number of occurrence in this corpus normalized by the total number of tokens in the same corpus N:

$$P(c) = \frac{Count(c)}{N} \qquad (4)$$

To avoid zero probabilities we use the add-one smoothing technique:

$$P(c) = \frac{Count(c) + 1}{N + V} \qquad (5)$$

where V is the size of the vocabulary of our corpus.

P(o/c) is the probability that the word c is misspelled as o, this probability depends on who the typist is familiar with the keyboard. Therefore, this probability cannot be computed exactly.

To get over this problem, we use the technique used by (Kernighan *et al*., 1990), and create a confusion matrix for each editing operation. These matrices represent the number of times one letter was incorrectly used instead of another:

Del[x,y] the number of times xy was typed as x.

Ins[x,y] the number of times x was typed as xy.

Sub[x,y] the number of times x was typed as y.

Trans[x,y] the number of times xy was typed as yx.

Using these matrices one can estimate P(o/c) as :

$$P(o/c) = \begin{cases} \dfrac{del[c_{i-1}, c_i]}{count[c_{i-1}c_i]} \\ \dfrac{ins[c_{i-1}, o_i]}{count[c_{i-1}]} \\ \dfrac{sub[o_{i-1}, c_i]}{count[c_i]} \\ \dfrac{trans[c_i, c_{i+1}]}{count[c_i c_{i+1}]} \end{cases} \qquad (6)$$

where *i* is the transformation position in the word c to get the error o.

2.3. Noisy channel algorithm

1. Count the number of tokens N and the vocabulary size V in the corpus.
2. If a given word o is not in the vocabulary:

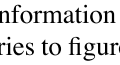
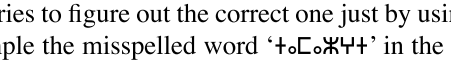
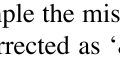
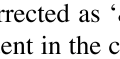
    a- From o, generate all possible words using a single deletion, insertion, substitution or transposition.

    b- The set C is made up of generated words belonging to the vocabulary

    c- For each word *c* in C, compute *P(o/c)* using (6)

    d- The proposed correction of the word *o* is given by (3)

3. Bigram language model

The Noisy channel algorithm failed to return the appropriate word because it does not use any information about the other words in a sentence. It deals only with the misspelled word and tries to figure out the correct one just by using single word frequencies in the corpus. For example the misspelled word '  ' in the sentence '  ' is corrected as '  ' instade of '  ', just because the first correct word is more frequent in the corpus than the seconde one. And the insertion of ' ' is more frequent than its deletion.

3.1. N-gram language model

The N-gram approach to spelling error detection and correction was proposed by Mays *et al*. (1991). In (Jurafsky et Martin, 2000) The idea is to generate every possible misspelling of each word in a sentence either just by typographical modifications (letter insertion, deletion, substitution, transposition), or by including homophones as well, (and presumably including the correct spelling), and then choosing the spelling that gives the sentence the highest prior probability. That is, given a sentence w = $\{w_1, w_2, ..., w_k, ..., w_n\}$, where w_k has alternative spelling w'_k, w''_k etc, we choose the spelling among these possible spellings that maximizes P(W).

In a general way, the probability of a sentence (sequence of words) is given using the chain rul as folow :

$$P(w_1, w_2, ..., w_n) = P(w_1^n) = \prod_{k=1}^{n} P(w_k / w_1^{k-1}) \qquad (7)$$

The terms $P(w_k / w_1^{k-1})$ are approximated by using Markov assumption:

$$P(w_k / w_1^{k-1}) \approx P(w_k / w_{k-N+1}^{k-1}) \qquad (8)$$

The N-gram model approximates the probability of a word given all the previous words $P(w_k / w_1^{k-1})$ by the conditional probability of the N-1 preceding words $P(w_k / w_{k-N}^{k-1})$.

3.2. Bigram language model

In a particular case, the bigram language model (N=2) assigns probability to sentences (string of words: $w_1, w_2, ..., w_n$) whether for computing probability of a sentence or for probabilistic prediction of the next word in a piece of a sentence as follows:

$$P(w_1, w_2, ..., w_n) = P(w_1^n) = \prod_{k=1}^{n} P(w_k / w_{k-1}) \qquad (9)$$

For example:

P(␣ا+◦⊙ ا +⊏⚡◦◩◩◦ا+) = P(␣ا+◦⊙/<s>)

x P(ا/␣ا+◦⊙)

x P(+⊏⚡◦◩◩◦ا+/ا)

The <s> is used to indicate the beginning of the sentence.

4. Algorithm

1. Count the number of tokens N and the vocabulary size V in the corpus.
2. For a given sentence S
3. For each word o in S
4. If o is not in the vocabulary:

    a- From o, generate all possible words using a single deletion, insertion, substitution or transposition.

    b- The set C is made up of generated words belonging to the vocabulary

    c- For each word c in C

        Compute *P(o/c)* using (6)

        Replace o by c in S and compute P(S) using (9)

        Score = *P(o/c)* x *P(S)*

    e- The proposed correction of the word o is the word c with the highest score.

5. Application and results

The most influent thing in this work is the corpus used to compute all probabilities. If the corpus used in the training stage is too specific to a domain, the probabilities will not generalize well the new test sentences and vice versa.

As a beginning, the corpus used is extracted from the IRCAM website. This corpus has N=3322 tokens of a vocabulary size V=893. Algorithms are tested using a Java program. As shows the figure 1, the algorithm has detected the misspelled word 'ⵜⴰⵛⵣⵢⵜ' and proposed the right correction 'ⵜⴰⵛⴰⵣⵢⵜ'.

The accuracy is about 32%. This is due to the small corpus and the language model used (Bigram) that can't model langue sentences.

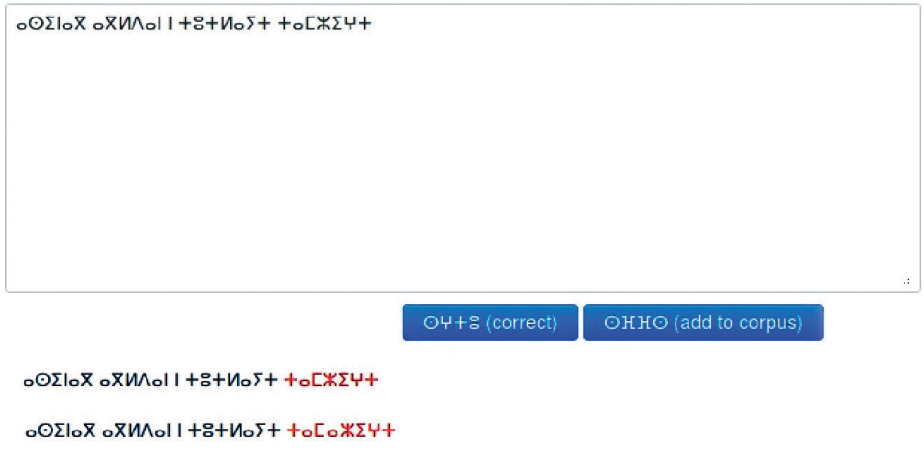


Figure 1: An example of the system application

6. Conclusion

In this work, we have presented the automatic spelling correction applied to Tamazight language written in Tifinagh. We used the noisy channel algorithm and the bigram language model. The most important issue is the corpus design.

References

Jurafsky D., Martin H. (2000). *Speech and language processing*. Prentice Hall.

Kernighan M. D., Church K. W., Gale W. A. (1990). A Spelling Correction Program Based on a Noisy Channel Model. *AT&T Bell Laboratories*.

Réalisation d'un Editeur de Texte pour la Langue Amazighe et d'une Barre d'Outils Amazighe pour MS Office

Nourdine Ait Zengui[1] Ali Rachidi[2] El Houssine Bouyakhf[1]
[1]Université Mohamed V-Agdal, Faculté des Sciences de Rabat
Laboratoire d'Informatique Mathématiques Appliquées Intelligence Artificielle
et Reconnaissance de Formes
bouyakhf@mtds.com ait.zengui@gmail.com
[2]Ecole Nationale de Commerce et de Gestion, B.P. 37/S Hay Salam – Agadir, Maroc
Laboratoire de Traitement d'Images et Systèmes d'Information
rachidi.ali@menara.ma

1. Introduction

L'utilisation des technologies nouvelles de l'information et de communication (TIC) est devenue essentielle pour assurer la survie des patrimoines dans la société. En effet, l'élaboration des outils de traitement et d'analyse de texte de l'amazighe peuvent apporter un considérable appui pour que la langue amazighe et de réelles chances de survivre sur le web et dans le monde informatisé.

L'application de ces nouvelles technologies permettra de mettre en évidence les phénomènes linguistiques pour l'amazighe. En effet, on pourra faire des traitements de texte et de courriers électroniques ou même effectuer des recherches ou exploiter les fonctions syntaxiques et grammaticale de cette langue qui est bien peu doté informatiquement. Par conséquent, des recherches scientifiques et linguistiques sont lancées pour remédier à cette situation. L'un des volets prioritaire de cette recherche, est de concevoir et de réaliser des applications capables de traiter de façon automatique des données linguistiques Amazighes.

Le présent papier présente un système d'écriture ou un éditeur de texte indépendant et autonome qui traite le texte amazighe. Il présente aussi une barre d'outils qui se charge automatiquement dans les applications de MS Office qui configure le système au traitement de texte amazighe.

2. Amazighe : langue naturelle

2.1. Présentation

La langue amazighe est la langue la plus anciennement attestée au Maghreb. Son aire couvre prés de cinq million de km^2, elle s'étend d'est en ouest de la frontière égypto-libyenne aux Iles Canaries, et du nord au sud de la rive méridionale de la Méditerranée au Niger, Mali et au Burkina Faso. La communauté la plus importante dont l'amazighe est la langue première se trouve au Maroc. De par son antériorité, la langue amazighe constitue le mode

d'expression de l'identité première des Marocains; elle représente un fondement essentiel de leur environnement socioculturel comme elle façonne leur inconscient collectif et marque leur personnalité de base. Elle joue présentement le rôle de creuset dans la formation du mouvement culturel amazighe.

2.2. L'alphabet Amazighe

L'IRCAM a proposé à l'Organisation de Standarisation Internationnale (21/06/2004) l'Alphabet Tifinaghe et ce dernier a été confirmé. Cette proposition comprend quatre sous-ensembles de caractères tifinaghes :

1. le jeu de base de l'IRCAM ;
2. le jeu étendu de l'IRCAM ;
3. d'autres lettres néotifinaghes en usage ;
4. des lettres touarègues modernes dont l'usage est attesté.

La liste de l'alphabet Tifinaghe et le plan Unicode associé attribué par l'ISO est illustré à la figure 1.

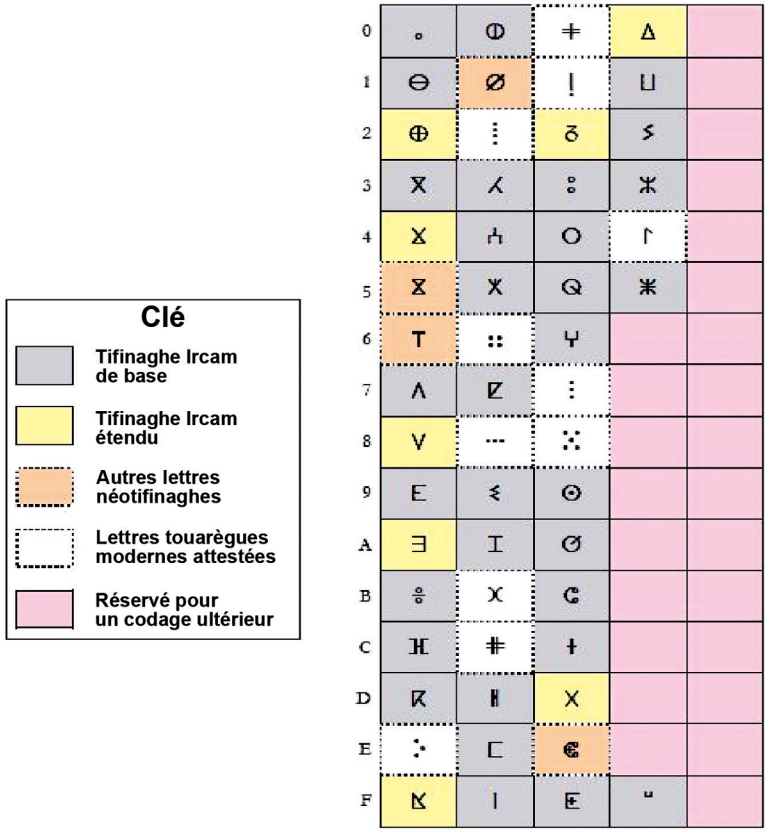


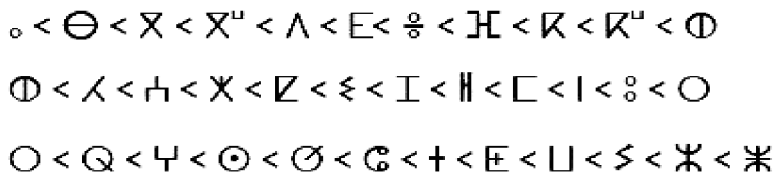
Figure 1 : Plan unicode associé au Tifinaghe

2.3. Ponctuation

Nous ne connaissons pas de signe de ponctuation particulier au tifinaghe.L'IRCAM a préconisé l'emploi des signes conventionnels qu'on retrouve dans les écritures latines : « » (espace), « . », « , », « ; », « : », « ? », « ! », « … », Etc. En conséquence, cette proposition ne présente aucun signe de ponctuation Tifinaghe.

2.4. Tri

Seul l'IRCAM a défini un ordre précis décrit par l'expression ci-dessous (a < b, signifie que a est trié avant b) :



2.5. Chiffre

L'IRCAM a retenu les chiffres « arabes » occidentaux (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) pour l'écriture Tifinaghe. Cette proposition n'introduit donc aucun nouveau chiffre ou nombre.

2.6. Directionnalité

L'IRCAM a retenu la direction horizontale de gauche à droite pour l'écriture Tifinaghe.

2.7. Clavier Amazighe

L'alphabet Tifinaghe est composé de trente trois caractères. La figure 2 illustre un clavier sous format ASCII proposé par le centre CEISIC de l'IRCAM[1].

1 Institut Royal de la Culture Amazighe, Centre des Etudes Informatiques et des Systèmes d'Information et de Communication :

- L. Zenkouar, Y. Belkasmi, Y. Aït Ouguengay, *Élaboration d'une première version du clavier amazighe*, Plan d'action 2003. *cf*. aussi http://www.IRCAM.ma/Telecharger/pilote.html.
- *Conception et mise au point des polices ci-dessous*, Plan d'action 2003. *cf*. http://www.IRCAM.ma/Telecharger/polices.html.
  - Police standard Tifinaghe-Ircam, L. Zenkouar, Y. Aït Ouguengay, H. Jaa;
  - Police Tifinaghe-Ircam izzuren, Y. Aït Ouguengay, H. Aarab, L. Zenkouar;
  - Police Tifinaghe-Ircam taromit, Y. Aït Ouguengay, H. Aarab, L. Zenkouar;
  - Police Tifinaghe-Ircam tissnat n'irrumin, Y. Aït Ouguengay, H. Aarab, L. Zenkouar;

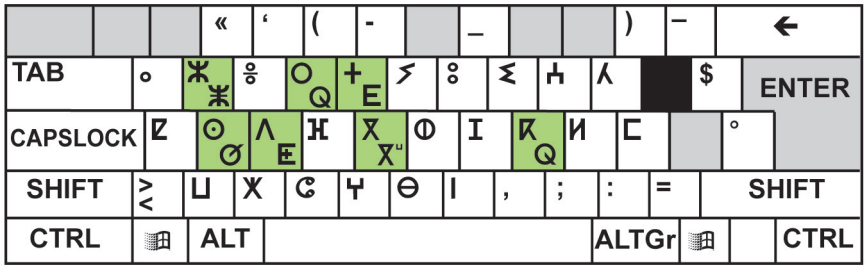


Figure 2 : Clavier Tifinaghe sous format ASCII. Les 26 premiers caractères sont accessibles directement. Les caractères emphases s'obtiennent en utilisant la case Noir (le « ^ » en clavier latin) de la même façon qu'on utilise le « ^ » en français (pour taper le « â »)

3. Conception d'un éditeur indépendant et autonome qui traite le texte amazighe

3.1. Présentation

La langue amazighe est peu dotée informatiquement. En effet, la réalisation d'un éditeur de texte indépendant et autonome pour le traitement du texte amazighe est considéré comme une étape importante pour l'informatisation de l'amazighe. La première partie de notre travail est de réaliser un éditeur de texte de la langue amazighe.

L'idée de concevoir un éditeur de texte pour la langue amazighe est venue du fait que les éditeurs de texte présents sur le marché du traitement de texte n'offrent pas toutes les fonctionnalités nécessaires pour un traitement de texte de qualité.

L'éditeur de texte qu'on essaye de réaliser permet d'assurer les fonctionnalités de base du traitement de texte telles que :

- Ouvrir un fichier
- Saisie du texte amazighe
- Mise en forme canonique du texte sélectionné (standardisation, saisie non univoque)
- Copier-coller du texte amazighe
- Changement de police
- Sauvegarder le fichier sous format RTF

La fonctionnalité qu'on traitera de plus pour cet éditeur et la fonction de recherche et du remplacement du texte amazighe, car cette opération n'est pas prise en compte pour la langue amazighe sur les éditeurs de texte déjà présents.

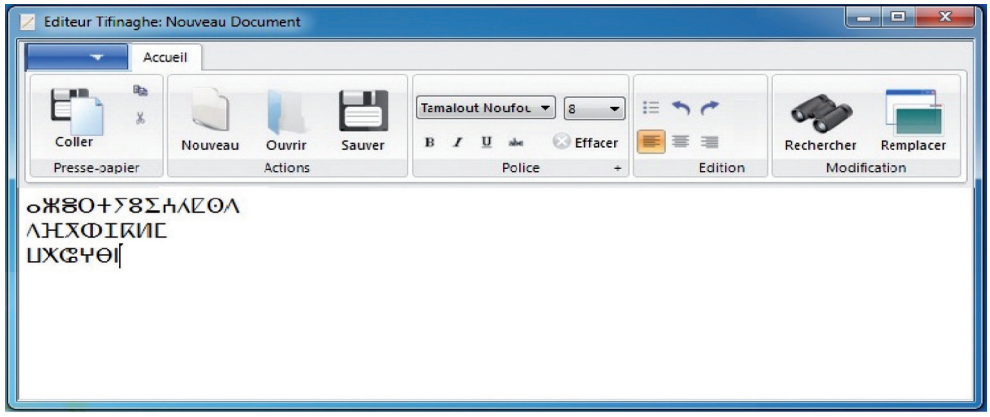


Figure 3 : Editeur de texte de la langue amazighe

3.2. Mise en œuvre et outils impliqués

Pour l'élaboration de l'éditeur de la langue amazighe, on a utilisé le langage de programmation C#. Car c'est un langage orienté objet à typage fort créé par Microsoft. Il a été créé afin que la plate-forme Microsoft .NET soit dotée d'un langage permettant d'utiliser toutes ses capacités. Il est très proche du Java dont il reprend la syntaxe générale ainsi que les concepts.

Le choix du langage de programmation est dû au fait que les chaînes de caractères dans C# sont en Unicode, ce qui facilitera l'intégration des polices et caractères amazighes élaboré déjà par l'IRCAM.

Le clavier et les polices amazighes sont déjà élaborés par le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication à l'IRCAM pour tous les systèmes d'exploitation Windows et pour Mac.

La fonction ajoutée par rapport aux éditeurs de texte développés jusqu'à présent pour le traitement de la langue amazighe est la fonction de recherche et de remplacement d'une chaîne de caractères amazighes. En effet, notre éditeur possède bien cette fonctionnalité de plus ce qui lui permet de faire le travail des éditeurs les plus connues comme MS Office.

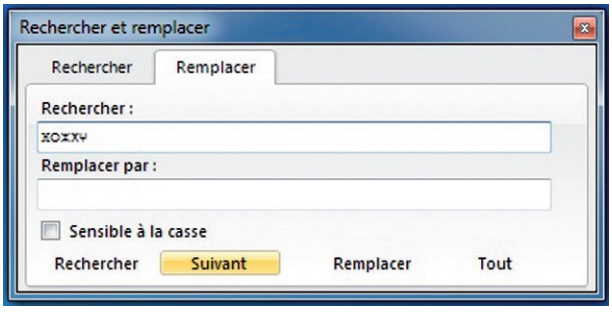


Figure 4 : Fenêtre de recherche et de remplacement du texte amazighe

4. Réalisation d'une barre d'outils Amazighe pour MS Office

4.1. Présentation

Microsoft Office est une suite bureautique éditée par Microsoft pour les plates-formes Windows et Macintosh. Cette suite inclut plusieurs logiciels dont Microsoft Word qui est un logiciel de traitement de texte qui est considéré comme le programme central de Microsoft Office. Il domine les marchés du logiciel du traitement de texte.

L'intégration de la langue amazighe à Microsoft Office est un point en faveur de la langue. Car ça permettra un traitement fiable et de qualité du texte amazighe. Pour cela la deuxième partie de notre travail est de réaliser une barre d'outils pour Microsoft Office qui configure le système au traitement du texte amazighe.

L'objectif de ce travail est de réaliser comme première étape ce kit (barre d'outils) qui se charge parmi les autres barres d'outils de l'office qui permettra de réaliser cette configuration.

4.2. Mise en œuvre et outils impliqués

Pour la réalisation de la barre d'outils de MS Office, la solution choisie est d'utiliser l'environnement de travail Microsoft Visual Studio 2010 intègré de nouvelles fonctionnalités qui simplifient le processus de développement d'application. Parmi ces fonctionnalités, on trouve le développement des Add-in de MS Office.

Un Add-in est une partie complémentaire d'un logiciel de base. Il peut être sous la forme d'un menu principal, un menu contextuel ou même une fonction d'arrière plan. Ici nous allons nous intéresser au développement d'un menu en vue de faire intégrer les fonctionnalités qui manquent pour le traitement du texte par la suite bureautique MS Office.

Les fonctionnalités qu'on cherche à intégrer au MS Office sont principalement le tri, la recherche et le remplacement d'un mot ou d'un texte amazighe, et avoir l'accès à un dictionnaire de la langue amazighe.

Le menu qu'on nommera AMAZIGH sera présenté comme suit :

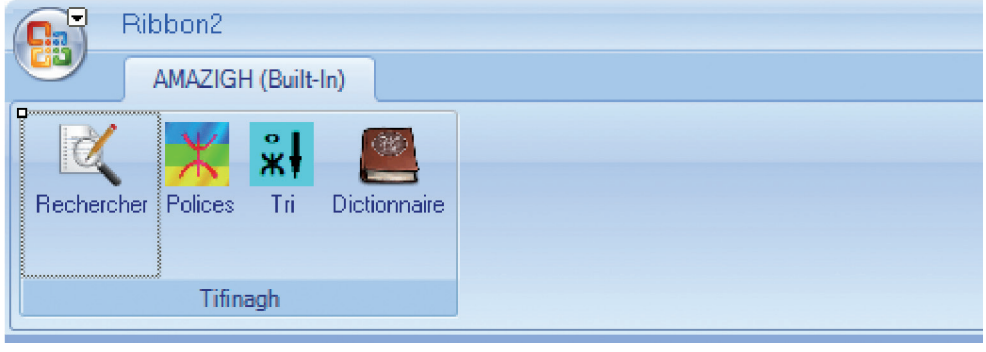


Figure 5 : Le menu AMAZIGH

De la gauche vers la droite, les fonctions des boutons sont :

- Le bouton Rechercher qui permet de faire une recherche ou un remplacement d'un texte amazighe.
- Le bouton Polices qui permet de choisir une police amazighe.
- Le bouton Tri donne accès au tri des tableaux en amazighe. Cette fonction permet d'ordonner les tableaux selon un ordre lexicographique.
- Le bouton Dictionnaire donne accès à un dictionnaire amazighe-français.

Le développement des Add-in se fait par l'un des langages de programmation Visual Basic (VBA) ou même C#.

Pour ce travail on est arrivé à l'étape finale qui est de faire fonctionner cet Add-in. Mais la seule contrainte qui se pose c'est le dictionnaire amazighe car pour le moment il n'y a que des dictionnaires sous forme de documents en format PDF, ou même un dictionnaire en ligne sur internet comme par exemple celui qui est présent sur le site internet http://www.amazighnews.com/dictionnaire/.

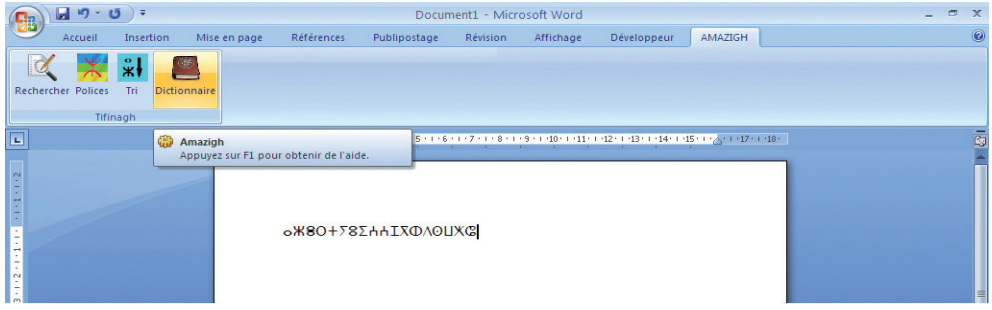


Figure 6: L'ADD-in AMAZIGH en execution dans MS Word

Le travail est toujours en cours sur ce projet afin d'améliorer la suite MS Office pour le traitement de texte amazighe.

4. Conclusion

L'informatisation de la langue amazighe est devenue un facteur déterminant de son usage et de sa survie. C'est un processus qui demande du temps et des efforts et qui ne peut se réaliser que de manière progressive. C'est pourquoi qu'une telle tâche ne pourra se réaliser qu'en adoptant des projets qui peuvent assurer l'utilisation de la langue amazighe d'une manière simple et accessible à tous le monde.

Références

Rachidi A., Mammass D. (2005). Informatisation de La Langue Amazighe : Méthodes et Mises En Œuvre, *3rd International Conference: Sciences of Electronic Technologies of Information and Telecommunications*, March, 27-31,Tunisia.

Berment V. (2004). Méthodes pour informatiser des langues et des groupes de langues « peu dotées ». *Thèse de Doctorat de l'université Joseph Fourier*, Grenoble 1.

ANMorph: Amazigh Nouns Morphological Analyzer

Hanae Raiss **Violetta Cavalli-Sforza**
Al Akhawayn University
raiss.hanae@gmail.com V.CavalliSforza@aui.ma

Abstract

We present ANMorph, a morphological analyzer/generator for Amazigh nouns, based on the Stuttgart Finite State Transducer platform. It was developed using the LCTL Berber v1.0 language pack, produced by the Linguistic Data Consortium in collaboration with IRCAM. ANMorph achieves over 90% correct results in the analysis of 1541 nouns extracted from the corpus. While these results are promising, more testing and evaluation are needed to enhance and refine coverage of nouns, and especially to extend the analyzer to other part-of-speech categories.

1. Introduction

Morphological analysis is a core part of several NLP applications, providing information about the possible part of speech and other morpho-syntactic features of words and tokens, as they appear in context, to the higher levels of linguistic analysis. Stems, suffixes, prefixes, and features such as tense, gender, and number are all elements that a morphological analyzer can extract depending on the language, the input being analyzed, and the needs of the user.

In the last several years, much effort has been invested in building corpora and tools for strategically and economically important languages such as Arabic, but significantly less attention has been paid to languages like Amazigh. Even in Morocco, where it is spoken by a significant portion of the population, the language has been neglected for decades. Though several linguistic studies exist, resources supporting computational treatment of the language, including corpora and electronic dictionaries, remain scarce relative to those for many other languages. Resource scarcity and the lack of orthographic standardization characteristic of mostly spoken languages have been obstacles to computational processing of Amazigh and development of automatic tools. Fortunately, interest in the language has increased in the last decade, especially since the creation of IRCAM (the Royal Institute of the Amazigh Culture) in 2001, with the objective of promoting the language at all levels (Boukhris *et al*., 2008). The position of the Amazigh language in Morocco was also reinforced by its integration

into the Moroccan educational system in 2003 and its establishment as one of the official languages of Morocco with its inclusion the Moroccan constitution in 2011[1]. The next step is strengthening its presence in the area of information technology and making the Amazigh language accessible to all through electronic resources and automatic processing tools.

This paper presents ANMorph, a two-level morphological analyzer for Amazigh nouns, implemented using the open-source Stuttgart Finite State Transducer (SFST) tools[2]. ANMorph is not complete: further development and evaluation must be performed to make it a complete morphological analyzer for Amazigh nouns and significant further work will be required to include other parts of speech, including verbs. However, even in its current state, ANMorph does achieve the objectives of building a corpus of Amazigh nouns and providing rules for analyzing a large part of them. For this reason, we consider it a contribution to the still scarce set of tools and resources available for processing this language.

2. Characteristics of the Amazigh Language

The Amazigh language, also called Berber, is an ancient language in the Afro-Asiatic or Hamito-Semitic language family. It is a morphologically rich language characterized by a large number of dialects due to historical, geographical, and sociolinguistic causes (Ameur *et al*., 2006). Geographically speaking, it spreads in "North Africa from the Siwa Oasis in Egypt in the east to Senegal in the west, and from Algeria in the north to Mali in the south" (Eifring and Theil, 2005), and was also found in the Canary Islands. Despite the fact that the Amazigh language is still spoken in several North African nations, it is not admitted as one of the official languages in any of the nations except Morocco. Among the reasons behind the existence of many Amazigh dialects is also the fact that the language remained mainly in the spoken sphere and was not regularized or centralized as one written language.

In Morocco, we distinguish three major Amazigh dialects (Eifring and Theil, 2005). Tarifit is spoken in northern Morocco, Tamazight in the Middle Atlas and south-eastern Morocco, and Tashelhit in south-western Morocco and the High Atlas (Ameur *et al*. 2010).

2.1. Amazigh Morphology

The morphology of a language describes the different surface representations that words can have. Amazigh morphology is rich in terms of its inflections and complexities it produces. It covers five main lexical categories: nouns, verbs, adjectives, adverbs, and prepositions (Sadiqi, 1997). Practically speaking, nouns and verbs are the base of the Amazigh morphology and the more important categories to focus on, as others can be derived from them. The focus of this work is noun morphology.

1 Title 1, Article 5 of the Moroccan Constitution (Secretariat of the Government, 2011)
2 http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

2.2. Noun Characteristics

The noun in the Amazigh language is always composed of one word between two spaces and does not start with an article (but see Oulhaj, 2000). It is characterized by gender (masculine or feminine), number (singular or plural), and state (free or construct/annexed).

2.2.1. Gender

Masculine nouns start with 'a', 'i', or 'u' with some exceptions, which include kinship nouns. To transform a singular masculine noun to feminine,'t' is added to the beginning and the end.

Example:

- *ifri* ⵉⴼⵔⵉ (cave, m.) —> *tifrit* ⵜⵉⴼⵔⵉⵜ (small cave, f.)

Unless the noun is inherently meaningful in both genders, the feminine of a noun that is originally masculine frequently has a diminutive meaning.

2.2.2. Number

Most masculine singular nouns in the Amazigh language start with 'a' (ⴰ); nouns that start with an 'i' (ⵉ) are next in frequency, followed by nouns starting with 'u' (ⵓ). The plural form is derived from the singular of the noun with the addition of some affixes and some vocalic changes. The latter affect the initial letters/sounds in the noun: the sound 'a' changes to 'i' or to 'u' (Sadiqi, 1997). To mark the plural, the sound 'an' is added to the end for a masculine noun, occasioning a vowel change if the singular noun ends in a vowel. For most feminine nouns, the 'a' after the initial 't' changes to 'i'; the final 't' drops and is replaced by 'in'.

Examples:

- *asrdun* ⴰⵙⵔⴷⵓⵏ (mule, m. sg.) —> *isrdunan* ⵉⵙⵔⴷⵓⵏⴰⵏ (mules, m.pl.)
- *ifri* ⵉⴼⵔⵉ (cave, m.sg.) —> *ifran* ⵉⴼⵔⴰⵏ (caves, m.pl.)
- *uccen*[3] ⵓⵛⵛⵏ (wolf, m.sg.) —> *uccenan* ⵓⵛⵛⵏⴰⵏ (wolves, m.pl.)
- *tamUart* ⵜⴰⵎⵓⴰⵔⵜ (woman, f. sg.) —> *timUarin* ⵜⵉⵎⵓⴰⵔⵉⵏ (women, f. pl.)

2.2.3. State

The nouns are known by two states: the free state and the construct or annexed state. The free state is the normal unmarked form of all nouns. Each state, in turn, is divided into two categories: abstract or concrete. The concrete is also categorized into animate and inanimate nouns. We say that a noun is in the construct state when there is a change at the level of the first vowel in the following contexts (Sadiqi and Ennaji, 2004):

- When the noun follows the verb in a sentence and is the subject of the verb;
- When the noun follows a preposition;
- After some numerals.

3 We use LDC l-gram representation for the Latin transliteration of Tifinagh; 'c' is the 'š' or 'sh' sound.

In the construct state, feminine nouns drop the ‘a’ after the initial ‘t’, but in masculine nouns:

- Initial ‘a’ (ⴰ) changes to ‘u’ (ⵓ);
- Initial ‘i’ (ⵉ) changes to ‘y’ (ⵢ); and
- Initial ‘u’ (ⵓ) changes to ‘wu’ (ⵡ) (Sadiqi, 1997).

2.3. Challenges in Amazigh NLP

Various factors contribute to making Amazigh a complex language to process. First, in spite of standardization efforts, dialects (Tarifit, Tamazight, Tashelhit) are still present in the resources available. A concrete example is the corpus on which this work is based (LCTL Berber v1.0, from the Linguistic Data Consortium, http://www.ldc.upenn.edu/), which contains texts from newspapers and from IRCAM, with spelling variations due to differences between dialects. Inconsistent orthography is also likely to be due to the fact that most speakers of Amazigh are not used to writing it and so spell the words as they hear them without following a convention. A second challenge in computational processing of Amazigh, as mentioned earlier, is the scarcity of standardized and annotated corpora, because of the delayed attention paid to this language. A third reason for processing complexity is contractions. For example, the kinship term *baba* ⴱⴰⴱⴰ (father) followed by a pronoun *inas* ⵉⵏⴰⵙ (his), becomes the single word *babas* ⴱⴰⴱⴰⵙ (his father) (Ameur *et al*., 2006; Bourkhis *et al*., 2008). Finally, like many other languages, Amazigh has part-of-speech ambiguity: a word might have more than one part of speech associated with it, depending on meaning and the context of its occurrence. For example, *igra* ⵉⴳⵔⴰ (step behind) is a verb in one context; in another context it can be the plural noun of *agru* ⴰⴳⵓ (frog) (Outahajala, 2011).

3. ANMorph: Development and Evaluation

3.1. The Stuttgart Finite State Transducer - SFST

To develop ANMorph, we used the Stuttgart Finite State Transducer (SFST) platform. SFST is a non-commercial, open-source, freely-available toolkit that supports the analysis, generation, and processing of finite-state automata and transducers. It was primarily developed to implement morphological analyzers. The tools run under Linux, are command-line oriented, and not so easy to use due to scarce documentation. The SFST toolset contains the following components, with programs implemented in C++ (Schmid, 2007):

- SFST-PL, a programming language for implementing finite state transducers (FSTs);
- A compiler for SFST tools, called **fst-compiler**;
- Tools for applying, comparing and printing FSTs, including **fst-mor**; and
- A FST library.

A program is first written using SFST-PL, which uses regular expressions and two-level rules. **fst-compiler** compiles the program into minimized FSTs, which are run by **fst-mor** to perform morphological analysis and generation (bidirectionality is an advantage of FSTs).

3.2. Data Description and Data Processing

The data we used to develop ANMorph was developed by LDC under the REFLEX Less Commonly Taught Languages (LCTL) program (http://projects.ldc.upenn.edu/LCTL/), whose objective was to create digital resources to support the less commonly taught languages, also called low density languages due to the lack of resources. Amazigh was one of the languages chosen for inclusion (LDC, 2010; Cieri and Liberman, 2010). Language resources and tools are organized in a language pack. The one we received included documentation, some data (monolingual text, parallel text and named entity annotations), and a few of the tools (encoding converter, tokenizer, named entity tagger, word and sentence segmenter). Each type of data is divided into training (90%) and evaluation (10%) sets and provided in different formats: 'original_docs' (original documents in text format) and 'txt/' (UTF-8 text files), 'ltf/' (XML files containing the segmented UTF-8 texts). The work described in this project used primarily monolingual data (most of which is from Le Monde Amazigh newspaper), with the other resources, primarily parallel text, used as supplementary information to aid in understanding the monolingual text, but not actively exploited. We also received from LDC two lexicons developed outside the LCTL project, a word frequency list computed from text believed to have standardized orthography, and an initial set of morphology rules in a format not usable by SFST. We used this data, especially the lexicons and word frequency list, for comparison with and enrichment of the basic monolingual data in the LCTL language pack.

The monolingual training data contains 85 XML files of varying size. We processed 58 of these to: 1) extract individual tokens, 2) convert them to LDC-1 gram (see Table 1), and 3) identify and select nouns tokens. SFST requires as input a text file with one token per line.

IRCAM Tifinagh Alphabet	Equivalence in Arabic	Equivalence in Latin Alphabet	Example in Amazigh	LDC 1-gram	Example using LDC 1-gram
ⵔ	ر	ṛ	bṛṛa (outside)	R	bRRa
ⵟ	ط	ṭ	tiṭṭ (eye)	T	tiTT
ⵣ	ز	ẓ	iẓi (liver)	Z	iZi
ⴹ	ض	ḍ	aḍaṛ (foot)	E	aEaR
ⵃ	ح	ḥ	aḥidus (musical Amazigh group)	H	aHidus
ⵄ	ع	ɛ	aɛdis (stomach)	3	a3dis
ⵖ	غ	ɣ	aɣrum (bread)	U	aUrum
ⵚ	ص	ṣ	tamṣṛiyt (Egyptian)	S	tamSRiyt
ⴽⵯ	ك	k^w	amdkk^{w}l (friend)	K	amdkKl
ⵅⵯ	ک	g^w	azgg^{w}aɣ (red)	G	azgGaU

Table 1: Partial mapping between IRCAM Tifinagh alphabet and LDC 1-gram encoding

LDC 1-gram is a transliteration/encoding scheme that uses only ASCII characters and maps each Tifinagh character to a single LDC 1-gram character. We chose this encoding over Tifnagh because it is compatible with SFST and because of its simple relation to Tifinagh. Most of the data we received from LDC is written in a Latin-based encoding, but not one that

could be directly used by SFST. Hence, we converted it to LDC 1-gram using the encoding converter script contained in the Berber language pack.

3.3. Lexicon Final Results

From the 58 XML monolingual training files, we selected 1541 words recognizable as nouns in Tarifit (the mother tongue of one of the authors) from 13606 tokens in total, and categorized them into: Free nouns (1048), the normal noun category, such as *ifri* ⵉⴼⵔⵉ (cave) and *aman* ⴰⵎⴰⵏ (water); Proper nouns (91), such as *Anir* ⴰⵏⵉⵔ; Borrowed nouns (299), from Moroccan Darija or Standard Arabic, Spanish, or French; Kinship nouns (15), such as *baba* ⴱⴰⴱⴰ (father), *yema* ⵢⵜⵎⴰ (mother); and Non-inflected nouns (88), like *anZar* ⴰⵏⵣⴰⵔ (rain).

3.5. Implementation of ANMorph

To implement ANMorph using SFST several components and steps are went through. An example of each step is shown in Table 2 and further discussed below.

ANMorph Phases	Example
1. Reading the lexicon	$WORDS$= "nouns"
2. Defining the alphabet	ALPHABET= [A-Za-z]<B> <E> <noun><sg><plr><masc><fem>
3. Adding inflectional endings	$Nmasc-Infl$ = <noun> (<plr><masc> | <sg><masc>)
4. Adding Boundary Markers	S = <>:<B> S <>:<E>
5. Defining Inflection Rules	$R1$ = {<sg><masc>} :{} ^-> (<noun> __)
6. Identifying the Morphological Rules	$R5$ = ([aeiu]:<>) ^-> (__ <noun> an)
7. Deleting the Marker Symbols	[A-Za-z] [<B> <E> <noun>]:<>
8. Combining the rules	S = S || $R1$ || $R2$ || $R3$ || $R4$ || $R5$ || $R6$ S = S || $R7$ || $R8$ || $R9$ || $R10$ || $R11$ || $R12$
9. Compiling the SFST Morphological Analyzer	`fst-compiler morph.fst morph.a`
10. Running the SFST Morphological Analyzer	`fst-mor morph.a`

Table 2: SFST steps and sample rules for Amazigh nouns

Step 1: An example of a SFST entry in the lexicon is **izzri<Nmasc>** (song).

Step 2: The alphabet, containing the valid symbol pairs for the two-level rules, indicates that both lower and upper case characters are accepted. The markers **<B>**, **<E>**, **<noun>**, **<sg>**, **<plr>**, **<masc>**, and **<fem>** respectively mark the beginning character of a word the end character of a word, part of speech, gender and number. A replacement rule in SFST has the general form C ^ -> (L __ R) where C is the replacement to apply. The left side is always the *analysis layer*; the right is the *surface layer*. To restrict rule application, left and/or right context are specified with the general form (L __ R), where L and R are surface strings that must match in the surface layer. Two-level rules also use an intermediate layer to store intermediate symbols, like markers and variables, which are removed in the surface layer.

Step 3: Defines allowable inflectional endings for plural and singular masculine nouns stems.

Step 4: Adds boundary markers, which are used to mark the start and the end of interesting segments; for example, to add a specific character at the end of a stem, the marker <E> will be used to mark the end of the stem.

Step 5: Defines the inflection rules of each inflectional ending. Rule $R1$ means that, if a nouns is singular and masculine, no ending should be added to it (**:{ }**). It is a replacement that maps '**<noun><sg><masc>**' to '**<noun>**'.

Step 6: Specifies the morphological rules that describe the regular stem changes of a noun; for example the stem vowel changes in pluralization. Table 3 shows two of the morphological rules in ANMorph, providing an explanation and an example of each. Rule $R5$ is applied in the pluralization of nouns that end with 'i', 'a', and 'u' when 'an' (the plural suffix for masculine nouns) is added. Rule $R6$ replaces the 'a' with 'i' in nouns that starts with 'a' when the inflectional ending 'an' is present that indicate masculine plural nouns.

Step 7: The markers are deleted from the surface layer since, after applying the rules, the markers are no longer useful.

Step 8: The complete resulting transducer is created by conjoining all the rules. The order in which rules are given does not matter much, as they are combined at the end by conjunction.

Step 9: The analyzer, consisting of definitions and rules provided in steps 1-7 and assembled in Step 8, is stored in file 'morph.fst' and is compiled with **fst-compiler** into file 'morph.a'.

Step 10: The analyzer, file 'morph.a', is run with **fst-mor**; the user enters the words to analyze.

The rules in SFST-PL	Explanation	Examples
`$R5$ = ([aeiu]:<>) ^-> ( __ <noun> an)`	Deletes 'i', 'a', 'u' before the inflectional ending 'an'	ifri ⵉⴼⵔⵉ (cave) → ifran ⵉⴼⵔⴰⵏ
`$R6$ = a:i ^-> (<B>__ .* <noun>an)`	Replaces 'a' with 'i' only when plural inflection is present.	adar ⴰⴷⴰⵔ (foot) → idaran ⵉⴷⴰⵔⴰⵏ asrdun ⴰⵙⵔⴷⵓⵏ (monkey) → isrdunan ⵉⵙⵔⴷⵓⵏⴰⵏ

Table 3: Some morphological rules for Amazigh nouns

ANMorph also recognizes some context-dependent orthographic changes, for example when 'a' changes to 'u', or 'i' to 'y' or 'u' to 'w' as in 'azru' —> 'uzru'. To handle this, we define the alphabet with all symbols used in the surface layer of the transducer. Then, we add a marker **<prep>** to mark the orthographic changes. **uzru** is analyzed as **<prep>azru** where <prep> indicates that this form is only valid after a preposition and is meant to be part of the analysis

of the word ‘uzru’. This is summarized in the following syntax, added before Step 8:

ALPHABET = ^S	*Includes in alphabet all symbols from surface layer of transducer*
S= <prep>? S	Add the prepositional marker for context, ‘?’is the optionality operator
S=S ǁ (<prep>:<>a:u)?.*	Add the in-context transformation to transducer

3.6. Evaluation of Results

At this stage of its development, ANMorph contains a total of 12 rules, of which one addresses context changes. In terms of performance, out of the 1541 unique nouns analyzed by ANMorph, only 132 (9.32%) were not correctly analyzed. The incorrect analyses are mostly due to the fact that some of the nouns do not adhere to all the rules; for example the rule for conversion to feminine is applicable, but the one for pluralization is not. For example: *argan* ⵄⵔⴳⴰⵏ => targant ⵜⵄⵔⴳⴰⵏⵜ but not argan ⵄⵔⴳⴰⵏ => *arganin* ⵄⵔⴳⴰⵏⵉⵏ or *targant* ⵜⵄⵔⴳⴰⵏⵜ => tirganin ⵜⵉⵄⵔⴳⴰⵏⵉⵏ. There are also cases where rules are applied when they should not be. To avoid these problems, rules must be refined to cover only appropriate cases.

A remark regarding the issue of allomorphs, e.g. {‘azru’, ‘uzru’}, currently represented in our rule set by the in-context transformation mentioned earlier, is in order. Generally speaking, morphological analysis and its counterpart, generation, concern single words only. Analysis is used to extract and tag stems, as well as inflectional and derivational morphemes. For some languages, other prefixes and suffixes that are attached to the word, may also be separated out (e.g. prefixes like و or ف and direct object pronoun suffixes in Arabic). However, changes due to interaction between words, such as occur in Amazigh in the construct state, are not the domain of morphology per se. In machine translation, for example, they are treated in a pre-processing or post-processing step. Handling them in morphology may result in ambiguous analyses if one of the allomorphs of a word also occurs as a distinct word in the lexicon.

4. Summary and Future Work

It is only recently, through the work of IRCAM and other actors, that significant efforts have been made to restore the Amazigh language and culture, to integrate it within the educational structure of Morocco, and to give the language more visibility nationally and internationally through publications and participation in worldwide conferences. Overall, however, little attention has been devoted to date to building the resources and tools necessary for computational processing of the Amazigh language. Our motivation to work on a morphological analyzer for Amazigh came from this scarcity of computational infrastructure, a morphological analyzer being one of the fundamental tools in many NLP tasks.

We consider ANMorph just a start towards building a morphological analyzer for Amazigh. The choice to focus only on nouns at first resulted from the multiple simultaneous challenges encountered in using the SFST tools, which provide finite state technology for building two-level morphological analyzers but are not well documented or easy to use, the complexity of Amazigh morphology, and the not altogether standardized orthography of our basic data,

which also required some processing to clean and format for use with SFST. While our morphological rules do not capture all regularities even for nouns, we are encouraged by our results, which leave only slightly under 10% of the nouns improperly analyzed.

With regards to future work and in order for the system to have better coverage and be mature, we intend to address the following points:

- Provide improved coverage of irregularities, that is, inflection of nouns that deviate from the existing rules;
- Enlarge the lexicon to include nouns from the two other major dialects, Tamazight and Tashelhit, which we have already started to do;
- Include other part of speech in the morphological analyzer;

Check our lexicon against widely used Amazigh dictionaries (e.g. Chafik 1990, 1996, 1999, which was not available to us at the time of writing, and Ameur *et al*., 1996), for orthographic differences and to verify coverage of different classes of nouns.

In order for morphological rules to work correctly on all dialects, it may be necessary to add dialect tags to both lexicon and rules. Moreover, the rules should be tested on data subsets for specific dialects, as well as nouns shared by all dialects, to assess the sensitivity of rule performance to dialectal variation.

References

Aït Ouguengay Y., Bouhjar A. (2010). For standardised Amazigh linguistic resources. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC-2010, Malta, May 17-23, pp. 2699-2701.

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk , A., El Medlaoui, A., Iazzi, M., Souifi, H. (2010). مدخل إلى اللغة الأمازيغية [Introduction to the Amazigh language]. Rabat: Royal Institute of the Amazigh Culture.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., El Medlaoui M. and Iazzi E. (2006). *Graphie et orthographe de l'Amazighe*. Rabat: IRCAM.

Ameur M., Bouhjar A., El Medlaoui M., and M. Iazzi (2006). *Vocabulaire de la langue amazighe (Français – Amazighe)*. Rabat: El Maarif Al-Jadida.

Ataa Allah F., Jaa H. (2009). Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue amazighe. In *Actes du 1er Symposium International sur le Traitement Automatique de la Culture Amazighe (SITACAM 2009),* Agadir, Maroc, pp. 110-119.

Ataa Allah F., Boulaknadel S. (2010). Light morphology processing for Amazighe Language. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC-2010*.

Boukhris, F. Boumalk, A. El Moujahid, E., Souifi, H. (2008). *La nouvelle grammaire de l'Amazighe*. Rabat: IRCAM.

Chafik, Mohamed (1990, 1996, 1999). *Dictionnaire bilingue : arabe-amazigh* (tomes 1, 2, 3). Publications de l'Académie Marocaine.

Cieri, C., Liberman, M. (2008), *15 Years of Language Resource Creation and Sharing: a Progress Report on LDC Activities*. LREC, 2008, Retrieved May, 10, 2012, from: http://www.lrec-conf.org/proceedings/lrec2008/summaries/861.html

Eifring, H., Theil, R., (2005), *Linguistics for students of Asian and African Languages*. Blindern: Universitet I Oslo. Retrieved on, May, 19th, 2012 from: http://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/index.html

Oulhaj, L. (2000). Grammaire du Tamazight. Eléments pour une standardization. Centre Tarik ibn Zyad pour les études et la recherche.

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010). Tagging Amazighe with AncoraPipe. In *Proceedings of Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation (LREC-2010),* Malta, pp. 52-56.

Outahajala M., Zenkouar L., Rosso P. (2011a). Building an Annotated Corpus for Amazigh. In *Post-Proc. 4th Int. Conf. on Amazigh and ICT, NTIC-2011*, Rabat, Morocco.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2011b). POS Tagging in Amazigh using Tokenization and n-gram Character Feature Set. In *Proceedings of Symposium Int. sur le Traitement Automatique de la Culture Amazighe (SITACAM-2011),* Agadir, Morocco.

Sadiqi F., Ennaji, M. (2004). *A Grammar of Amazigh*. Fes: Faculty of Letters, Dhar El Mehraz.

Sadiqi F. (1997). *Grammaire du berbère*. Paris: L'Harmattan.

Schmid, H. (2007). *SFST Manual and SFST Tutorial*. Retrieved, June 15, 2011, from: http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

Simpson, H., Maeda, K., Cieri , C. (2009). Basic Language Resources for Diverse Asian Langauge: A streamlined Approach for Ressource Creation. In *Proceedings of the 7th Workshop on Asian Language Resources*, Suntec, Singapore, pp. 55–62.

الأمانة العامة للحكومة (Secretariat of the Government) (2011). *La constitution (The constitution)*. Retrieved, June, 6, 2011, from: http://www.sgg.gov.ma/constitution_2011_Fr.pdf

TICAM'12, p. 137-146

Caractérisation de Voyelles et d'Emphatiques Berbères en vue d'une Identification de Locuteurs dépendant de Texte

Fatma zohra Chelali Amar Djeradi Hocine Teffahi
Laboratoire communication parlée et traitement du signal
Faculté d'électronique et d'informatique
Université des sciences et de Technologie Houari Boumedienne, B.P. 32 El-Alia,
Bab-Ezzouar, 16111, Alger, Algérie
Chelali_zohra@yahoo.fr adjerdi05@yahoo.com hteffahi@yahoo.fr

Résumé

Plusieurs études dans le développement des systèmes de Traitement Automatique de la Parole ont été présentées pour beaucoup de langues, mais les particularités de la langue berbère en raison de ces traits acoustiques spécifiques comme l'emphase qui caractérise le consonantisme du berbère posent de nombreux problèmes aux constructeurs des systèmes de reconnaissance automatique de la parole (RAP). Cet article repose sur une étude acoustique de la durée des voyelles et des emphatiques berbères ainsi qu'une analyse formantique qui se base sur le calcul des formants par le codage prédictif linéaire (LPC)). Cette paramétrisation est aussi utilisée pour établir un système de reconnaissance de locuteurs par le maximum de vraisemblance pour les trois voyelles et les cinq emphatiques berbères. Les résultats obtenus sont satisfaisants.

1. Introduction

Le berbère appartient à la famille dite Chamito-sémitique ou Afro-asiatique. La zone d'origine de cette famille de langues est très controversée avec un berceau primitif au Moyen Orient (Syrie-Palestine ou péninsule arabe) ou en Afrique (Afrique de l'Est) ou Sahara central. La langue berbère est actuellement en usage dans neuf pays sur le continent Africain : Mauritanie, Maroc, Algérie, Tunisie, Libye, Égypte, Niger, Mali, Burkina-Fasso (Dugoujon, 2006).

L'importance de la présence berbère varie selon ces pays. Des îlots plus ou moins denses témoignent de la présence berbère en Tunisie et en Libye (Galand, 1988, Chaker, 1978). Le berbère est parlé par une petite communauté vivant dans l'oasis de Siwa en Egypte. A tous ces groupes, s'ajoute un autre groupe berbérophone important : les Touaregs établis au Sud de l'Algérie, au Mali, au Niger et au Burkina Faso. Le berbère est principalement parlé en Algérie et au Maroc. Il est difficile d'avancer des chiffres sur l'importance démographique des populations berbérophones. On estime généralement, qu'au Maroc, le berbère est parlé par 35% à 40% de la population. En Algérie, ce pourcentage se situe autour de 25% (Ridouane, 2003).

Nous nous intéressons dans cet article à deux aspects : le premier concerne la caractérisation des voyelles et des emphatiques de la langue berbère ; tandis que le deuxième aspect traite le développement d'un système de reconnaissance de locuteurs dépendant de texte (mots isolés contenant les voyelles et les emphatiques étudiés).

L'article est présenté comme suit : la deuxième section donne un aperçu du système phonologique berbère, un intérêt particulier est porté sur les voyelles et emphatiques berbères. La troisième section présente le banc expérimental réalisé au niveau du laboratoire communication parlée et traitement du signal et les conditions d'acquisition du corpus berbère ainsi que les traitements effectués. La quatrième section présente les résultats obtenus. Enfin une conclusion et des perspectives sont présentées en section 5.

2. Système phonologique kabyle

Jusqu'à maintenant aucune étude n'a été effectuée pour établir un système phonologique kabyle à partir de l'étude de tous les parlers, les systèmes existants ont été élaborés en étudiant des parlers à part, nous pouvons citer comme exemple le système établit par S. CHAKER dans son étude du parler d'Ait Iraten et celui élaboré par R. KAHLOUCHE dans son étude phonologique du parler de Makouda (Chaker, 1978; Gaci, 2011) .

Le kabyle, langue à vocalisme pauvre (3 voyelles – au même titre que l'arabe – et deux semi-voyelles), mais au consonantisme très riche (38 phonèmes, qui sont toutes des consonnes), s'écrit et se lit de gauche à droite. La graphie des mots, en caractères latins, très proche de la transcription phonétique, traduit fidèlement la prononciation, voyelles et géminées comprises.

Exemple :/illa/—> « illa » — > « il est », « il existe ». (Makhlouf *et al*., 2006)

Figure 1: Comparaison des systèmes vocaliques arabe, berbère et français

Au même titre que l'arabe le kabyle présente un système vocalique pauvre (trois voyelles), a, u, i mais un système consonantique très riche :

	Bilabiale	Labio-dentale	Dentale	Alvéolaire	Post-alvéolaire	Palatale	Vélaire	Uvulaire	Pharyngale	Glottale
Occlusive	B p		ṭ ḍ t d		d 3 t ʃ dz		K g			?
Nasale	M		θ ð̣	s z						
Fricative		f v	ð	ṣ é	ʃ 3		ΛY	χ β	ħ ʕ	h
Approxi			r ô			j	w			
Latéral approxi			L							

Tableau 1 : Le système consonantique berbère (kabyle) (Makhlouf et al., 2006)

L'une des caractéristiques partagée par les systèmes consonantiques de toutes les langues berbères est le fait que toutes les consonnes simples ont des correspondantes géminées.

Le kabyle, à l'instar de toutes les autres variantes du berbère (Tamazight), ne comporte que trois voyelles orales, plus une voyelle de lecture :

- « a » est moins ouverte qu'en français. Ex : Aman (l'eau).
- « i » se prononce entre le [i] et le [e] de français. Ex : Izimer (un agneau).
- « u » se prononce [u] comme le [ou] du français. Ex : Uzzu n tayri (le chagrin d'amour).

Pour le « e » [ə] est appelé Schwa ou « ilem » en kabyle, il se prononce à peu prés comme le [e] anglais « children ». Il n'est pas considéré comme une voyelle, c'est plutôt une sorte lubrifiant phonétique dont le rôle est de faciliter la prononciation.

Les emphatiques en écriture amazighe

Le système phonique de la langue amazighe contient des sons qui ne sont différenciés que par l'emphase. Du point de vue articulatoire (articulation des sons d'une langue), l'emphase correspond à une différence d'écartement des organes d'articulation, c'est ce que les linguistes appellent le degré d'aperture. Le volume d'air déplacé lors de l'articulation du son, est alors plus important dans le cas où le son est emphatique.

3. Corpus acoustique réalisé

La chaîne de mesure du signal acoustique est constituée d'un microphone dynamique placé à quelques cm du locuteur. Le signal est numérisé directement sur une carte son à une fréquence d'échantillonnage 16 Khz 16bits. Le sujet doit prononcer 10 fois chacun des mots à l'aide du logiciel wavesurfer dont les paramètres d'acquisition tel que la durée d'enregistrement, la fréquence d'échantillonnage ainsi que le nombre de bits sont fixés à l'avance. Un traitement préliminaire des signaux obtenus tel que l'annulation de la composante continue et la segmentation manuelle peut être aussi réalisé par le même logiciel.

Pour réaliser ce travail, Nous avons élaboré une base de donnée constituée d'un corpus de treize mots isolés berbéres, trois mots contenant les voyelles et deux mots pour chaque emphatique bérbere, prononcés par six locuteurs (3 masculin et 3 féminin). Chaque mot a été prononcé avec une vitesse moyenne en assurant une bonne articulation. La séquence a été répétée dix fois par tous les locuteurs. Le nombre totale des mots analysés est (6 locuteurs * 13 mots *10 répetitions), ce qui nous donne 780 fichiers à analyser.

Pour la construction de la base de données, il a été demandé aux six locuteurs de lire avec la méme vitesse et un débit de parole moyen et en évitant toute perturbation extérieure des mots contenants les mots choisis pour l'étude. Les phonémes berbéres consernés par cette études sont telles que :

Des voyelles tels que :

/a/ dans le mot /tamart/ , /u/ dans le mot /tamurt/ et /i / dans le mot /tanmirt/ .

Des emphatiques tels que:

/ص/ dans le mot / yessub, issidh/, /ض/ dans le mot /adhar, assmidh/.

/ر/ dans le mot / yerwa, aghrum /, /ز/ dans le mot /tizurin, tazalit/ et /ط/ dans le mot / attas, amchttouh /.

3.1. Variation de la durée syllabique

Le but de cette étude est d'identifier les effets du contexte sur la durée des phonémes. Pour chaque mot des enregistrements, nous avons commencé par extraire les voyelles et les consonnes emphatiques après segmentation afin de calculer la durée de chacun des segments.

La figure 2 montre quelques variabilités de la durée pour les emphatiques berbères.

Nous remarquons donc que la durée du signal de parole calculée varie légerement pendant les dix répétitions et diffèrt aussi d'un locuteur à un autre. Ceci exprime bien le phénoméne de variabilité intra-locuteur et inter-locuteur qu'on peut le définir comme suit : Une même personne ne prononcera jamais deux fois le même mot de la même manière, cette variabilité intra-locuteur peut causer l'échec de la reconnaissance si elle n'est pas prise en compte.

Les caractéristiques physiologiques, sociales ou environnementales propres à chaque locuteur (âge, sexe, accent, …) impliquent des différences entre les prononciations d'un même mot.

Au niveau phonétique, la durée est un paramètre très important, elle caractérise les voyelles et les consonnes emphatiques de la langue berbère.

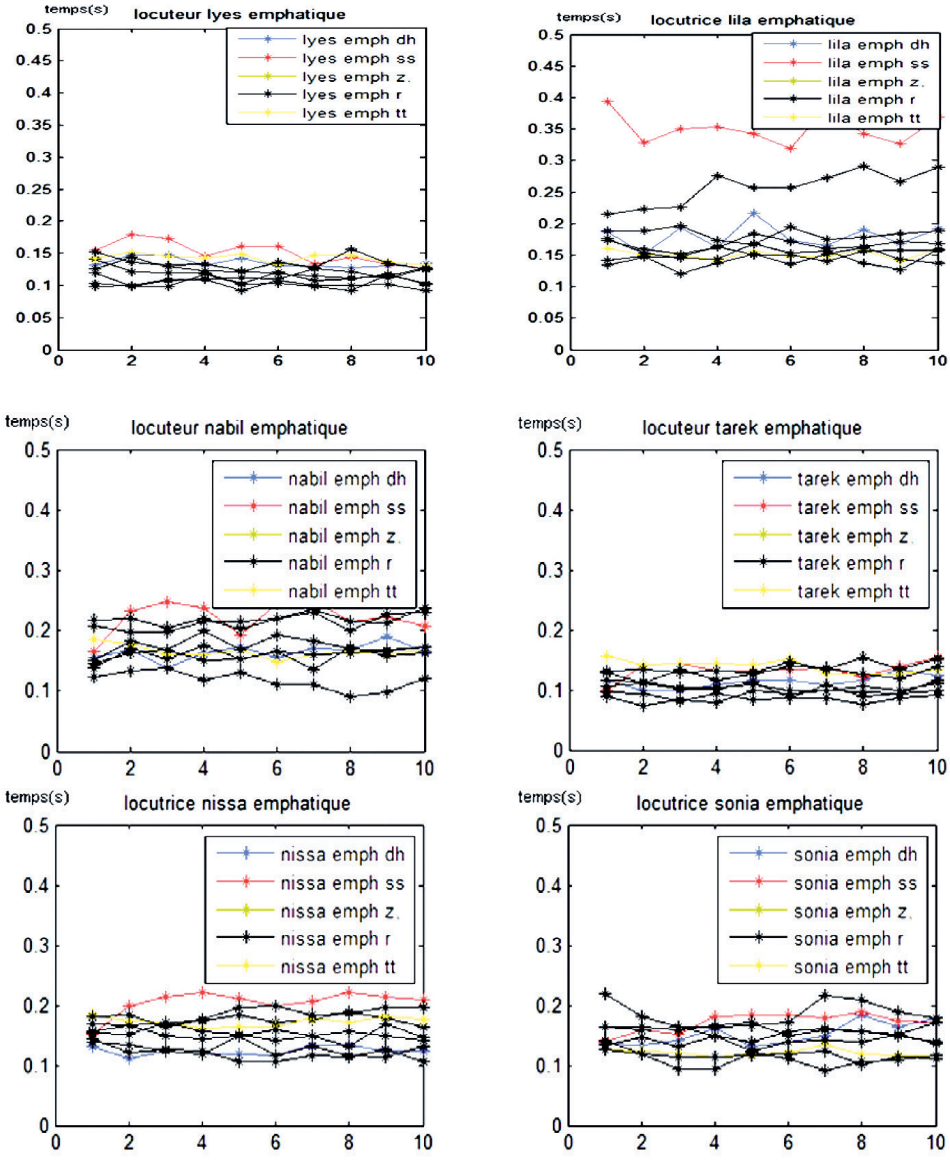


Figure 2 : Représentation des variations de la durée des emphatiques en fonction des répétitions

Comme les autres paramètres, la durée de l'entité choisie est largement dépendante du locuteur et du débit de parole. Ainsi, aucune mesure ne peut donner de modèle absolu de la durée. Comme nous l'avons déjà cité auparavant, la durée est corrélée à une multitude de facteurs complexes de nature linguistique (accent, position des mots dans la phrase, catégorie grammaticale, etc.) et extra-linguistique (débit de parole, expressivité, etc.).

3.2. Etude de la structure formantique

L'étude formantique est importante dans le traitement de la parole, elle nous permet de caractériser les voyelles et les consonnes. Pour les voyelles elles sont caractérisées par leurs premiers et le deuxième formant, alors que pour les consonnes il faudra rechercher le troisième et le quatrième formant jusqu'au cinquième.

Le calcul des formants a été fait par deux méthodes, la premiére est faite par programmation sous matlab v7.9 en se basant sur l'algorithme d'extraction automatique des formants par le codage prédictif linéaire LPC, et la deuxiéme est faite par le logiciel PRaat.

La figure 3 montre les variations des trois premiers formants (F1, F2, F3) pour les voyelles et la figure 4 pour les consonnes emphatiques.

- ***Cas des voyelles***

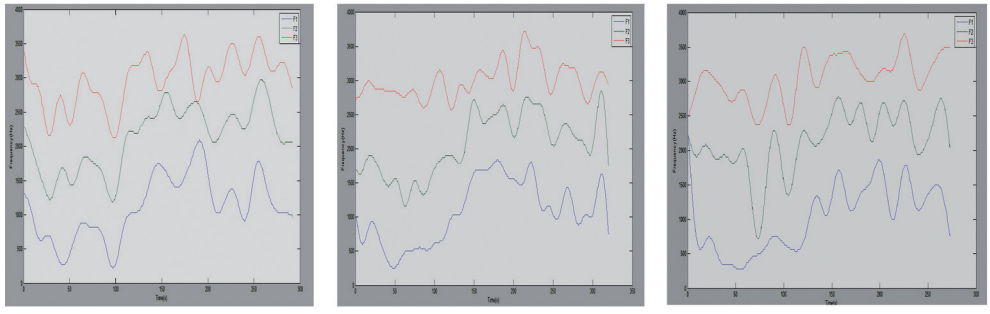


Figure 3 : Variation des formants F1 F2 F3 des voyelles (a u i) des mots 'tamart' pour les locuteurs 'nissa' 'nabil' 'sonia' respectivement.

- *Cas des emphatiques*

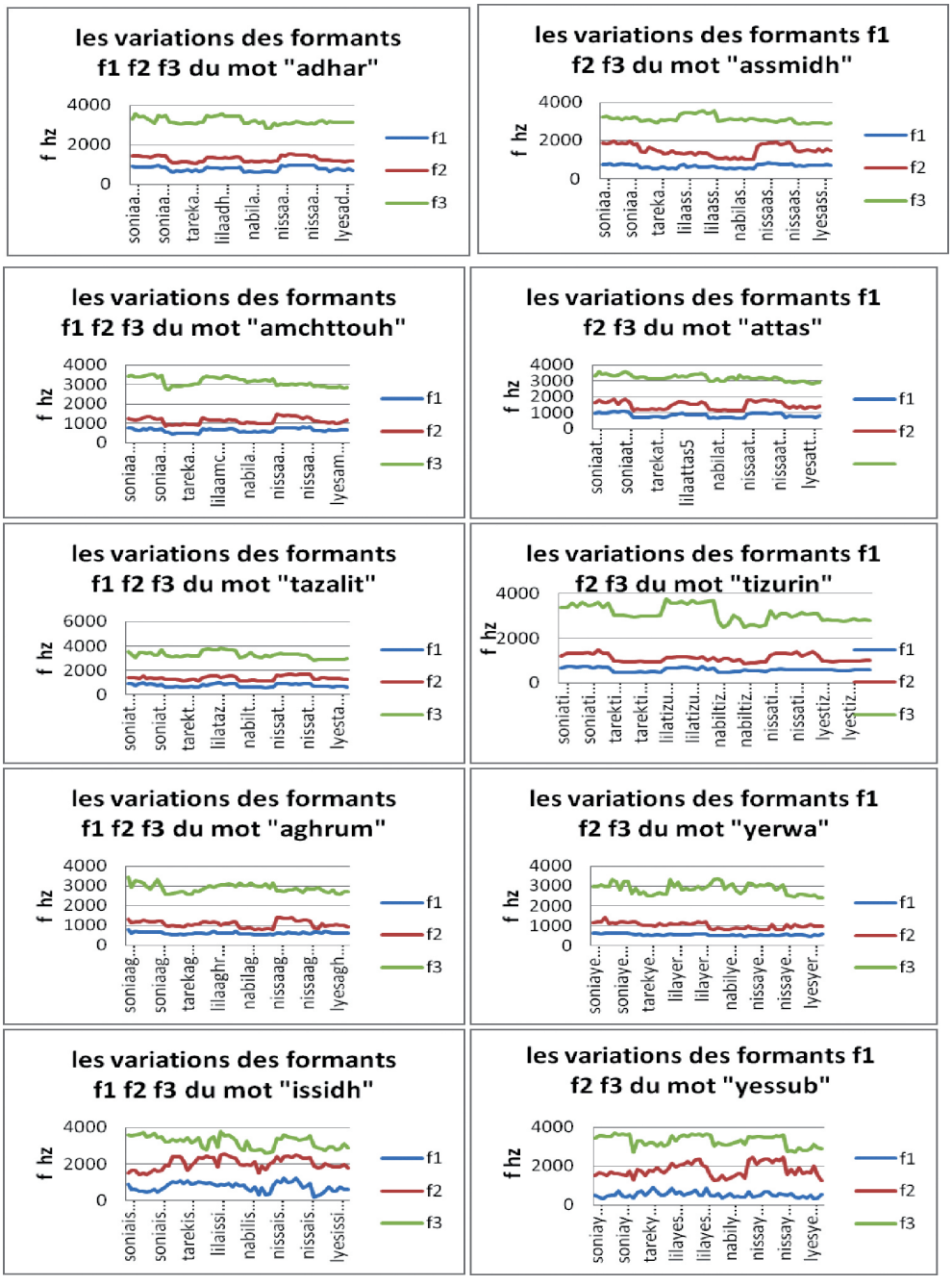


Figure 4 : Variation des formants pour les consonnes emphatiques

Nous remarquons donc que la variation des formants varie légerement pendant les dix répétitions et diffèrt aussi d'un locuteur à un autre et que les valeurs formantiques d'un locuteur féminin sont souvent élevées par rapport à un locuteur masculin.

- Pour les voyelles : F1 varie entre un minimum de 400 Hz et un maximum de 900 Hz, pour F2 entre 800 Hz et 2700 Hz, et F3 entre 2600 Hz et 3500 Hz .
- Pour les emphatiques : F1 varie entre 300 Hz et 900 Hz, pour F2 entre 1000 Hz et 2300 Hz, et F3 entre 2400 Hz et 3700 Hz.

En analysant les résultats obtenus, nous pouvons conclure :

Un abaissement des fréquences de premier formant et l'élévation de celle du deuxième constitue la caractéristique de l'emphase, l'élévation de F2 illustre le raccourcissement de la partie supérieure du conduit vocal. En revanche, l'abaissement de F1 correspond à l'augmentation de volume de la cavité antérieure du conduit vocal. Ceci prouve que le rétrécissement est localisé bas dans le pharynx. Ce qui confirme que la pharyngalisation est la définition articulatoire de l'emphase.

4. Système d'identification de locuteurs dépendant de mots berbères par analyse LPC

Il existe deux modes en reconnaissance automatique de locuteurs RAL : un mode dépendant du texte, où le texte prononcé par le locuteur est le même que celui utilisé lors de la session d'apprentissage, et le mode indépendant du texte, où le locuteur est libre de dicter un texte de son choix. En général les systèmes dépendants du texte donnent les meilleurs scores en reconnaissance.

Un système d'identification du locuteur se compose principalement de trois modules : paramétrisation, classification et décision. Le module de paramétrisation est un module d'extraction des paramètres pertinents. L'extraction de caractéristiques consiste à réduire l'information initialement présente dans le signal de parole et à le transformer en une séquence de vecteurs acoustiques robuste aux variations acoustiques. Nous avons choisi les vecteurs issus de l'analyse par prédiction linéaire LPC. A la sortie, le signal est représenté par un ensemble de vecteurs descripteurs appelés coefficients LPC calculés pour une largeur de fenêtre de Hamming de 256 points.

Le module de décision, qui selon les critères établis, permet de designer le locuteur reconnu. On utilisera dans ce travail la méthode de corrélation pour la reconnaissance.

Le classifieur réalisé est basé sur le critère de maximum de vraisemblance décrite par la méthode de corrélation. La corrélation consiste donc à rechercher le degré de ressemblance entre les deux vecteurs, c'est pourquoi elle repose sur le critère de vraisemblance.

$$r = \frac{S_{XY}}{S_{xx} S_{yy}} \qquad (1)$$

Où S_{XY} désigne la covariance de x et de y, définie par l'équation suivante :

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (2)$$

Et S_{xx} et S_{yy} désignent les écarts-type des distributions marginales en x et y. Le taux de reconnaissance de locuteurs est défini comme suit :

$$TR\% = \frac{Nombre\ de\ \ fichiers\ correctement\ classés}{nombre\ de\ \ fichiers\ test\ \ total} * 100 \qquad (3)$$

Le graphe montre le taux global de reconnaissance, il est pratiquement supérieur à 50% ce qui est acceptable, sauf, pour le dernier mot (yessub) pour l'emphatique /ص / qui est de 48.66%, cela est du à plusieurs paramètres tels que la variabilité intra et inter locuteur.

Nous pouvons conclure que le taux de reconnaissance varie pour chaque personne car la fréquence fondamentale de la voix est propre à chaque individu, elle est en fonction de différents paramétres physiologiques tels que le débit de parole, l'énergie, le volume, la masse de la glotte etc. D'autres classifieurs peuvent être envisagés tels que les classifieurs neuronaux et les machines à vecteurs de supports combinés aux vecteurs cepstraux MFCC ou PLP.

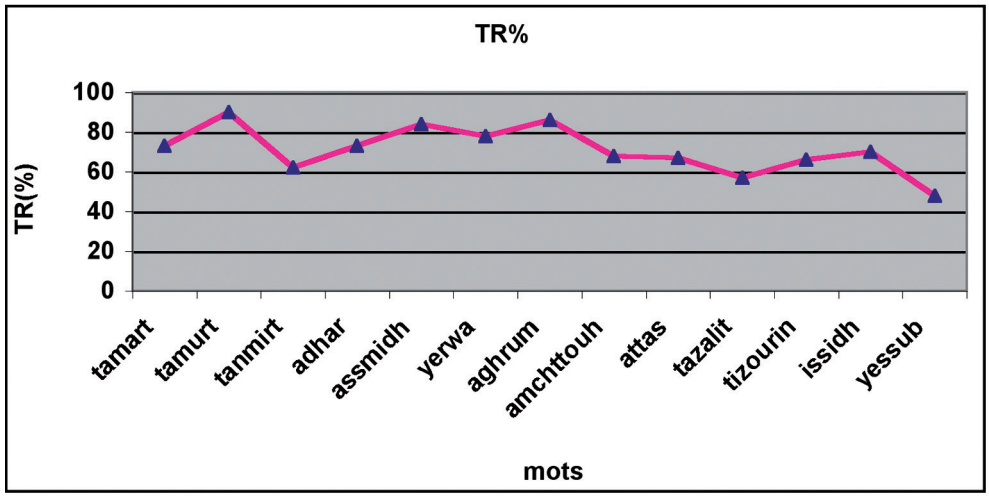


Figure 5 : Taux de reconnaissance de chaque mot pour les six locuteurs

5. Conclusion

Nous avons étudié dans une première étape les voyelles et les emphatiques de la langue berbère. Nous avons montré la variation de la durée syllabique des emphatiques d'un locuteur à un autre et pour le même locuteur. La durée est très importante dans l'étude acoustique de la langue berbère. Elle caractérise non seulement les voyelles, mais également les consonnes emphatiques. Nous avons montré aussi parmi les problèmes les plus importants liés à la reconnaissance automatique du locuteur la variabilité inter-locuteur et intra-locuteur dus à plusieurs facteurs cités précédemment. L'étude formantique nous a permis de dégager certaines caractéristiques de l'emphase en langue berbère.

La dernière partie de l'article introduit un système de reconnaissance de locuteurs par le critère de maximum de vraissemblance considéré comme linéaire. Les résultats sont satisfaisants et peuvent être améliorés en utilisant les coefficients cepstraux MFCC et les classifieurs non linéaires tels que les réseaux de neurones ou les machines à vecteurs supports SVM.

Références

Dugoujon Jean-Michel (2006). *Le berbère et les Berbères: Diversité linguistique et génétique*. Disponible sur: www.ohll.ish-lyon.cnrs.fr/pdf/dugoujon.pdf

Rachid Ridouane (2003). *Suites de consonnes chleuh: Phonétique et phonologie*. Université de la sorbonne nouvelle-paris III, Institut de linguistique et phonétique générales et appliquées.

Salem Chaker (1978). Un parler berbère d'Algérie (kabyle). *Thèse de doctorat d'état*, Paris V.

Med Makhlouf, Denis Legros, Brigitte Marin (2006). *Influence de la langue maternelle kabyle et arabe sur l'apprentissage de l'orthographe française*. http://www.cahiers-pedagogiques.com/IMG/pdf/Influence_langue_maternelle.pdf

Gaci Zohra (2011). Quel système d'écriture pour la langue berbère (le Kabyle)?. *Mémoire de magistère*, Université de Mouloud Mammri de Tizi Ouzou.

Teddy Dide (2007). *Réalisation d'un framework pour la reconnaissance de la parole*. Rapport de Stage, école polytechnique de l'université de tours.

Prosodie des Phrases Interrogatives et Affirmatives en Langue Berbère

Ramzi Halimouche Hocine Teffahi Leila Falek
L.C.P.T.S, Faculté d'Electronique et d'Informatique U.S.T.H.B, B.P. 32
El Alia Bab Ezzouar, Alger 16111, Algérie
rhalimouche@hotmail.com, hteffahi@gmail.com, lfalek@hotmail.com

Résumé

Ce travail s'inscrit dans un large mouvement international qui vise à ce que chaque peuple puisse disposer de tous les moyens pour communiquer dans sa langue. L'informatisation occupe ainsi une place essentielle dans cette vaste mobilisation culturelle et linguistique. Les TIC (technologies de l'information et de la communication) sont alors une des clés pour la sauvegarde, la diffusion, la connaissance et la visibilité des langues peu dotées.

Dans ce travail, nous allons préciser les différences prosodiques entre les phrases interrogatives et les phrases affirmatives en langue Kabyle. Pour cela, nous avons construit un corpus spécifique constitué des paires de phrases qui se composent chacune d'une phrase question et d'une phrase non-question. Les paramètres mesurés et analysés prennent en compte l'évolution de la prosodie, fréquence fondamentale et énergie, pendant l'énoncé de la phrase. Nous obtenons un taux de plus de 80% de bonne classification à l'aide d'un SVM bi-class.

1. Introduction

Utiliser la prosodie dans les systèmes de reconnaissance de la parole est un objectif de longue date, mais qui n'est toujours pas atteint. La prosodie pourrait être définie comme la mélodie du langage. Elle concerne toutes les variations de rythme et d'intonation dans les phrases. Les indices prosodiques sont directement accessibles depuis le signal acoustique, et correspondent à l'ensemble des modifications de l'énergie : de l'intensité liée à la force de la voix, de la fréquence fondamentale et la durée des différentes unités linguistiques (segments, syllabes). Les mots et les syllabes sont ainsi regroupés en unités prosodiques : et même si l'on ne connaît pas ce terme, tout le monde a l'intuition que, dans les phrases, on ne trouve pas des suites de mots isolés mais plutôt des groupes de mots les uns à la suite des autres. Ces groupes de mots sont les unités prosodiques, et celles-ci peuvent être intégrées dans une structure prosodique hiérarchique (Hirst *et al*., 2000).

Chaque langue possède un ensemble de contours intonatifs qui lui est propre. Ces contours sont liés aux types de phrases et à l'expressivité. A l'intérieur d'un même type de phrase, l'utilisation d'une intonation différente engendre des interprétations différentes. En effet, l'intonation est une des particularités prosodiques qui, pour une même phrase, donnent des sens (valeurs) divers (Vu, 2005).

Jusqu'à ce jour, très peu d'études ont analysé la prosodie de la langue kabyle en profondeur. Nous pouvons citer quelques travaux récents portant sur les tons lexicaux et sur la prosodie de la phrase (Chaker, 1991 ; Chaker, 1995 ; Chalah, 2007 ; Halimouche *et al*., 2012a). La question alors posée peut être résumée ainsi : existe-t-il, pour le berbère, langue à tons dont la prosodie est complexe, des informations extralinguistiques caractérisant le type de phrases, véhiculées par la prosodie et utilisées pendant les actes de dialogue pour la classification de ces types de phrases ? La réponse, outre le fait qu'elle nous permettra d'approfondir nos connaissances de la langue, si elle est positive, nous permettra d'envisager la réalisation de classifieurs automatiques indépendants des moteurs de reconnaissance.

2. Vecteur caractéristique

Le choix de l'élément prosodique approprié pour la constitution de vecteur caractéristique est justifié par un certain nombre de raisons. Premièrement, certains actes de dialogue sont ambigus. Par exemple, une question déclarative, « idhul wabridh ? », présente le même ordre de mots que la phrase affirmative correspondante « idhul wabridh » et, par conséquent, la prosodie est le seul moyen de pouvoir les distinguer. Deuxièmement, dans une application, l'exactitude de la reconnaissance du mot peut ne pas être parfaite. En effet, les systèmes de reconnaissance récents présentent encore un taux d'erreur de reconnaissance des mots de 30% pour le discours conversationnel. Par conséquent, le fait de compter sur des identités de mot peut propager systématiquement des erreurs du système de reconnaissance de la parole aux systèmes de traitement ultérieurs qui se basent sur le texte. Troisièmement, il y a des applications potentielles pour lesquelles on ne peut pas avoir une véritable reconnaissance de la parole disponible, et où la tâche consiste plutôt à dépister sommairement ce qui se produit dans un dialogue.

L'énergie est unique et bien définie pour toute portion de signal et aussi considéré comme le paramètre prosodique le moins complexe à calculer. Elle est d'ailleurs déjà utilisée couramment dans les systèmes de reconnaissance. L'énergie d'un signal échantillonné (xt) est définie par :

$$E = \sum \mathrm{x}(t)^2 \qquad (1)$$

Etant donné sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{dB} = 10.\log(\sum [X(t)]^2) \qquad (2)$$

Le calcul de l'énergie se fait sur des fenêtres courtes. L'ensemble des valeurs de toutes les fenêtres constitue le vecteur caractéristique.

Dans ce travail, nous allons utiliser deux paramètres prosodiques pour la définition du vecteur caractéristique : La mélodie, qui représente l'évolution de la fréquence fondamentale et la courbe d'évolution de l'intensité (énergie du signal).

3. Analyse du contour intonatif

3.1. Méthodologie et préparation du corpus

Partant de ces constats, nous souhaitons préciser les différences prosodiques entre les phrases interrogatives et les phrases affirmatives en langue berbère. Pour cela, nous avons construit un corpus spécifique constitué des paires de phrases qui se composent chacune d'une phrase question et d'une phrase non-question. Dans notre étude, nous avons porté une attention toute particulière au naturel des phrases, pour les phrases non-questions comme pour les phrases questions. Pour cela, nous les avons extraites de corpus reproduisant des situations de la vie courante. Quand il y a des particules dans la phrase interrogative, nous avons, dans la mesure du possible, gardé aussi les mêmes mots, ou bien nous avons utilisé des mots à la prononciation peu différente afin que ces deux phrases soient les plus ressemblantes que possible au niveau de la prononciation. De cette façon, nous éliminons ainsi tous les phénomènes de co-articulation qui pourraient interférer avec notre analyse prosodique. Le fait de choisir les mêmes tons nous permet d'éliminer l'influence des tons des syllabes sur l'intonation générale de la phrase. Toutes ces phrases ont été intégrées dans des dialogues significatifs, afin que leur prononciation soit la plus naturelle possible. Chaque dialogue est répété cinq fois par six locuteurs (4 hommes et 2 femmes) originaires de la Kabylie d'Algérie. Après sélection, notre corpus contient en tout, 720 phrases. Les paires de phrases utilisées sont présentées dans le tableau 1.

3.2. Résultats d'analyse

Pour chaque enregistrement nous analysons le contour de la fréquence fondamentale F_0 avec le logiciel Praat (exemple présenté figure 1).

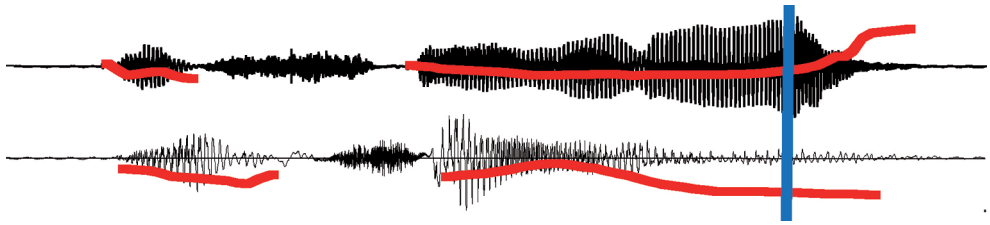


Figure 1 : Deux phrases à nombre de syllabes et tons identiques. En rouge le contour de F_0

En étudiant chaque paire de phrases présentées dans le tableau 1, nous remarquons que l'essentiel des différences d'intonation se situe à la fin de la phrase (zone située figure 1 après la barre verticale) : le contour de la dernière syllabe ou de la deuxième moitié de celle-ci semble être croissant pour les phrases interrogatives. Une étude statistique, présentée dans le tableau 2, confirme cette tendance : 70% des cas de phrases du type interrogative possèdent un contour de F_0 croissant à la fin de phrase et 97% des phrases de type affirmative possèdent un contour décroissant. Nous retrouvons là une tendance bien connue pour les langues non tonales comme le français (Rossi, 1999 ; Vu *et al*., 2005, Halimouche *et al*., 2011).

1.i	da cu dsa ?	*quel jour sommes nous ?*
1.a	asa d lkmis	*aujourd'hui c'est le jeudi*
2.i	atct agrum ?	*tu manges du pain ?*
2.a	ad tcag agrum	*je mange du pain*
3.i	mnhu i d kci ?	*qui es tu ?*
3.a	nki d lhusin	*je suis Hocine*
4.i	achal id yusan ?	*combien sont venus ?*
4.a	u sand atas	*beaucoup sont venus*
5.i	milmi ardyugal ?	*il revient quand ?*
5.a	adyugal azka	*il revient demain*
6.i	iwacu it atlat ?	*pourquoi t'as tardé ?*
6.a	idul wbrid	*la route est longue*
7.i	da cu itsbit ?	*qu'as-tu ramené ?*
7.a	abigd irdan	*j'ai ramené du blé*
8.i	achal wigi ?	*combien ceux là ?*
8.a	wigi sidrimn	*avec de l'argent*
9.i	idul wbrid ?	*le chemin est long ?*
9.a	abrid idul	*le chemin est long*
10.i	mnhu gkhm ?	*qui est à la maison ?*
10.a	akhm ytcur	*la maison est pleine*
11.i	usand inabgawan ?	*les invités sont venus ?*
11.a	An am usand	*oui ils sont venus*
12.i	anwa wigi ?	*qui sont ceux là ?*
12.a	wigi di mazigan	*ceux là sont des amazigh*

Tableau 1 : Les paires de phrases affirmatives (a) et interrogatives (i) du corpus

	Interrogative	*Affirmative*
Pente montante	*504 (70%)*	*22 (3%)*
Pente descendante	*216 (30%)*	*698 (97%)*

Tableau 2 : Nombre (et pourcentage) des contours de F_0 de la dernière moitié de la dernière syllabe

L'étude statistique, présentée dans la figure 2, donne le taux de détection correcte pour chaque locuteur suivant les contours de F_0 de la dernière moitié de la dernière syllabe. Nous pouvons affirmer que les paramètres prosodiques de la phrase berbère transportent des informations extralinguistiques qui peuvent permettre de discriminer le type de phrase. Comme pour les langues non tonales, ces informations sont essentiellement codées par le fait que l'intonation monte ou non en fin de phrase.

Au niveau production, cette première étape de l'étude a permis de caractériser la prosodie des phrases simples de la langue berbère (dialogue), en éliminant l'influence des tons: les différences entre questions et affirmations sont essentiellement une différence de pente de F_0 (croissante ou décroissante) en fin de la phrase (deuxième moitié de la dernière syllabe).

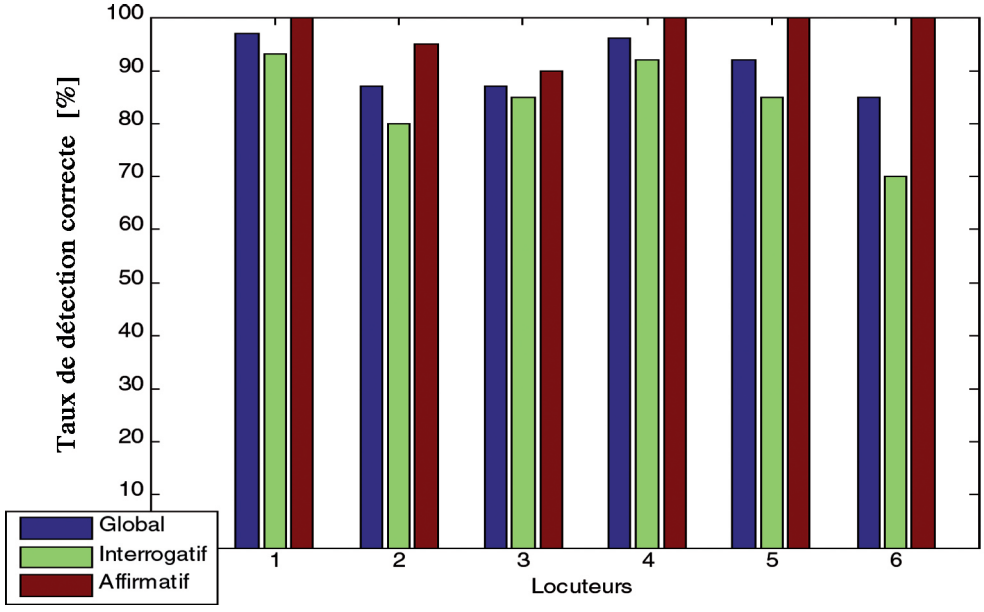


Figure 2 : Taux de détection correcte pour chaque locuteur suivant les contours de F_0 de la dernière moitié de la dernière syllabe

4. Classification automatique

4.1. Méthode utilisée

Pour évaluer automatiquement le type des phrases à des fins de classification en interrogatives et affirmatives, nous présentons dans ce paragraphe une méthode utilisant le contour prosodique. Les paramètres mesurés et analysés prennent en compte l'évolution de l'intonation pendant l'énoncé de la phrase, à savoir respectivement : augmentation des valeurs de F_0 en fin de phrase et évolution de l'énergie du signal.

Pour chaque enregistrement nous avons extrait le contour de la fréquence fondamentale F_0 avec une méthode de détection de F_0 basée sur la méthode cepstrale et l'évolution de l'intensité du signal. Nous calculons l'énergie de chaque phrase à court terme par la multiplication du signal par une fenêtre glissante de 10 échantillons, nous avons fixé la langueur de la fenêtre d'une façon expérimentale en faisant un compromis entre le taux de bonne identification et le temps d'exécution (Halimouche *et al*., 2012b).

Pour classifier nos phrases, nous avons adopté le formalisme des méthodes de Support Vecteur Machine (SVM) comme outil de décision discriminant. Le vecteur caractéristique, qui représente l'information pertinente, est la fréquence fondamentale puis l'évolution de l'énergie transportée par le signal.

Les Machines à Vecteurs de Support ou Séparateur à Vaste Marge (SVM) sont des techniques discriminantes dans la théorie de l'apprentissage statistique.

Elles ont été proposées en 1995 par V. Vapnik (Platt, 1998). Elles permettent d'aborder plusieurs problèmes divers et variés comme la régression, la classification, la fusion etc...

Nous avons appliqué dans notre travail un SVM mono-classe sous Matlab. Sachant que les SVM nécessitent une base de données d'apprentissage (2/3 de corpus) et une autre pour le test (1/3 de corpus) avec des vecteurs caractéristiques de même taille. Dans notre cas le corpus utilisé est constitué avec des phrases de différentes tailles ce qui donne des vecteurs caractéristiques différents, donc il est quasiment impossible de les utiliser dans un moteur de classification à base de SVM.

Pour s'affranchir de ce problème, nous avons opté pour une méthode de limitation des vecteurs caractéristiques (même taille) avec un taux maximum d'identification. Puisque l'information caractérisant le type se trouve à la fin de la phrase, le balayage commence de la fin du signal arrivant jusqu'au début du signal le plus court ; cela en calculant à chaque longueur du vecteur caractéristique le taux d'identification.

Nos expériences, sur la base de données précédente, ont montrées que le taux d'identification atteint la valeur maximum pour un vecteur caractéristique de 32 valeurs de (F_0) (maximum de l'information significative) et 29 valeurs pour l'énergie du signal. Au-delà de ces valeurs les taux de classifications se dégradent d'une manière significative (Halimouche *et al*., 2011).

Nous avons remarqué pendant nos analyses que le corpus contient du bruit au début et à la fin de chaque phrase dû à l'enregistrement (environnement bruité). Pour diminuer l'effet de ce bruit nous avons utilisé une méthode qui consiste à détecter les frontières des phrases et des

silences. Cette méthode est basée essentiellement sur l'estimation de l'énergie. Un silence ou bruit est détecté si l'énergie des 30 échantillons du début et de la fin du signal est inférieure à 10% de l'énergie maximum de la phrase. Cette opération permet de mieux estimer le contour intonatif (Halimouche *et al*., 2011; Halimouche *et al*., 2012b).

4.2. Résultats de la classification

La figure 3 donne le taux de classification correcte pour chaque locuteur suivant les contours de F_0. Nous remarquons que les phrases du type interrogatif sont les mieux identifiées pour les locuteurs (1, 2, 4, 5 et 6) par rapport aux phrases de type affirmatif. La figure montre que le locuteur 3 présente le taux le plus élevé de bonne identification globale. Ceci est dû essentiellement à un bon positionnement de ces articulateurs (notons, que les locuteurs ne sont pas issus de la même région de Kabylie). Le taux d'identification globale est de 86%. Dans l'ensemble, les phrases interrogatives sont identifiées avec un taux de 90%, pour un taux de 81% pour les phrases affirmatives (Halimouche *et al*., 2012a).

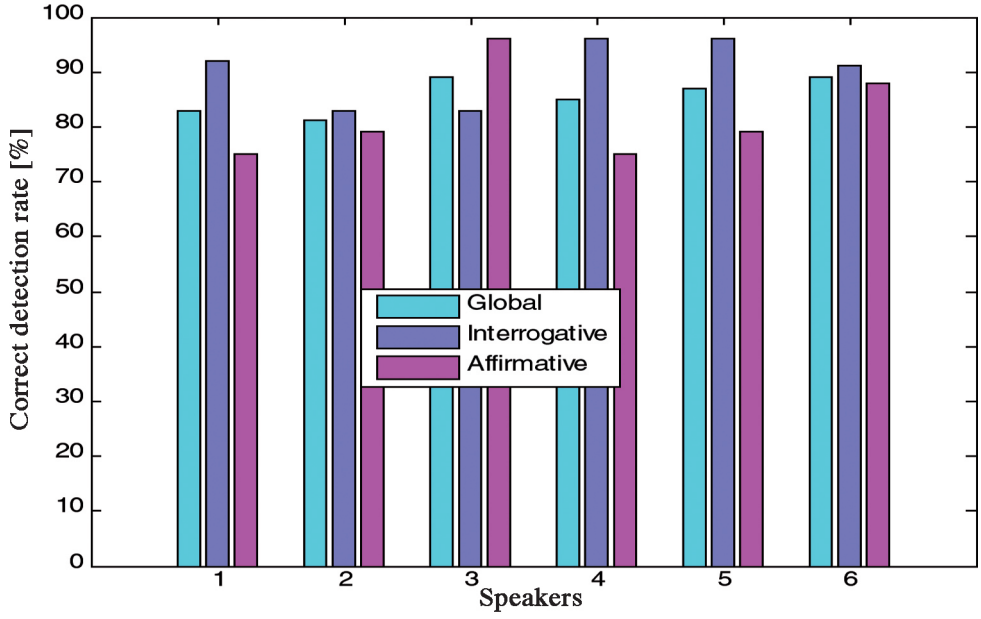


Figure 3: Taux de détection correcte pour chaque locuteur par un classifieur SVM
Vecteur caractéristique: fréquence fondamentale F_0

La figure 4 donne le taux de classification correcte pour chaque locuteur pour un vecteur caractéristique utilisant l'intensité (énergie du signal). Le taux d'identification globale est de l'ordre de 82%. Dans l'ensemble, les phrases interrogatives sont identifiées avec un taux de 85%, pour un taux de 78% pour les phrases affirmatives. Plusieurs travaux confirment que les phrases du type interrogatif sont prononcées avec un registre plus haut, ce qui fait

la distinction entre type de phrase affirmative et interrogative et confirme bien nos résultats (Halimouche *et al*., 2012b).

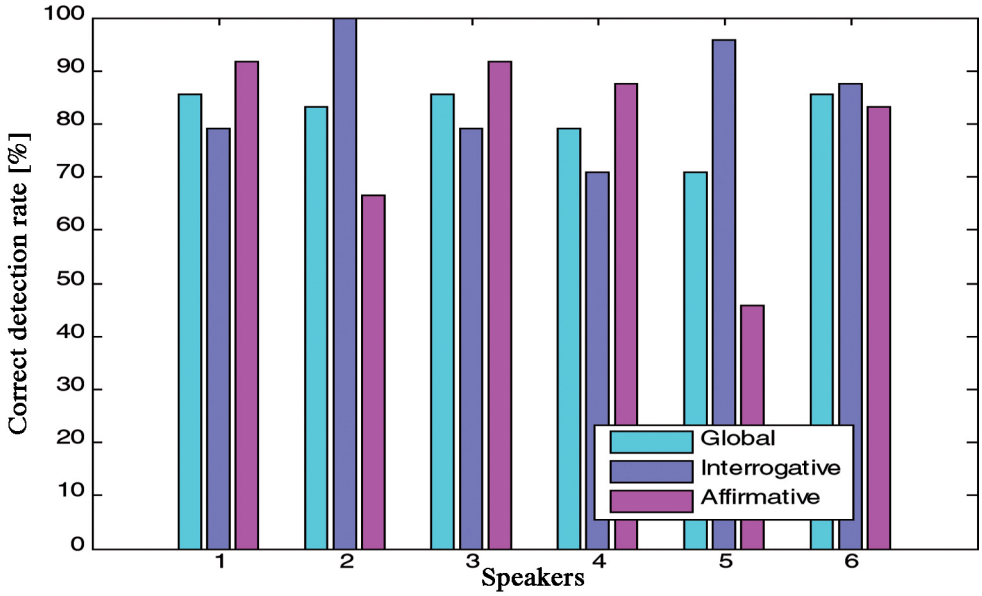


Figure 4 : Taux de détection correcte pour chaque locuteur par un classifieur SVM Vecteur caractéristique : énergie du signal

5. Conclusion

Dans ce travail, nous avons abordé le problème de détermination automatique du type de phrase pour la langue berbère. Nous avons présenté une nouvelle méthode de détection de questions par l'utilisation des SVM. Les vecteurs caractéristiques utilisés comprennent des paramètres prosodiques (F_0, intensité) qui sont extraits directement à partir du signal de parole.

Pour la langue berbère, nos recherches ont contribué à mieux comprendre la prosodie des phrases questions et nonquestions en berbère. Au niveau production, notre étude a permis de caractériser la prosodie des phrases simples de la langue berbère (dialogue), en éliminant l'influence des tons : les différences entre questions et nonquestions sont codées essentiellement par deux facteurs principaux : une différence de pente de F_0 (croissante ou décroissante) en fin de phrase (deuxième moitié de la dernière syllabe) et une différence du niveau d'intensité de la phrase. Nous avons construit une base de données contenant 720 phrases interrogatives et affirmatives et un modèle de classification fondé sur la prosodie avec un taux de classification correcte supérieur à 80%.

Les résultats de nos recherches pourraient être utilisés dans d'autres applications en parole telles que le résumé automatique, la navigation ou la recherche d'information, car les zones

autour d'une question contiennent souvent des informations importantes à identifier. De plus, les résultats de l'analyse de la production de la prosodie du berbère se révèlent fort utiles pour les applications de synthèse de parole de la langue berbère.

Références

Chaker S. (1995). Données exploratoires en prosodie berbère : l'accent kabyle. *Parues dans les Comptes rendus du GLECS*, 31, pp. 27-54,

Chaker S. (1991). Eléments de prosodie berbère : quelques données exploratoires. *Etudes et documents berbère*, 8, pp. 5-25.

Chalah C. (2007). Le rôle de l'intonation en syntaxe et en sémantique : étude de cas portant sur l'opposition d'état du nom kabyle. *Cahiers de l'ISL*, 22, pp. 47-62.

Halimouche R., Teffahi H. (2011). Vers une méthode de classification automatique de phrases interrogatives et affirmatives. *International conference on signal, image and their applications, SIVA'11*, pp. 145-148, Guelma, Algeria.

Halimouche R., Teffahi H. (2012). Intonation des phrases interrogatives et affirmatives en langue berbère. *Acoustics 2012*, Nantes, France.

Halimouche R., Teffahi H. (2012). Exploitation de l'Energie pour la classification automatique de phrases interrogatives et affirmatives. *International Conference on Multimedia Information Processing, CITIM 2012*, Mascara, Algeria.

Hirst H., Di Cristo A. (2000). Intonation Systems. A Survey of 20 Languages. Cambridge University, Press. Verlag.

Platt J. (1998). Fast Training of Support Vector Machines Using Sequential Minimal Optimization, in *Advances in Kernel Methods - Support Vector Learning*. eds. Schoelkopf B., Burges C., Smola A., MIT Press.

Rossi M. (1999). L'intonation, le système du français: description et modélisation. Ed Ophrys.

Shriberg E., Bates R., Taylor P., Stolcke A., Jurafsky D., Ries K., Cocarro N., Martin N., Meteer M., Van Ess-Dykema C. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech* 41, pp. 439-487.

Vu M., Castelli E., Boucher A., Besacier L. (2005). Classification de parole en Question et Non-Question par arbre de décision. SFC 05, *12ème Rencontres de la Société Francophone de Classification* - Montréal.

Effect of Dialect, Size of Population and PCA on Speaker Verification Performance

Kawthar Yasmine Zergat Abderrahmane Amrouche Nassim Asbai
Speech Com. & Signal Proc Lab-LCPTS
Faculty of Electronics and Computer Sciences, USTHB, Bab Ezzouar, 16 111, Algeria
{kzergat, namrouche, nasbai}@usthb.dz

Résumé

This paper investigates on the size of database and dialect of speakers which speak the same language for automatic text-independent speaker verification task. Statistical methods: Gaussian Mixture Model (GMM) and Support Vector Machines (SVM) were combined to obtain the hybrid GMM-SVM model.

Features vectors are extracted from the TIMIT corpus. They are represented by the MFCC (Mel frequency cepstral coefficients) and their first and second derivatives. In fact, the aim of this work is to apply the Principal Component Analysis (PCA) in front-end part in order to reduce the high dimensionality required for training the feature and speed up the training process. Performance evaluations show that transforming the data into a lower dimensional space using PCA improves significantly the recognition accuracy. On the other hand, the size of database has more influence than the dialect in automatic speaker verification task.

1. Introduction

Automatic speaker verification is an important embranchment of speaker recognition (Reynolds, 2002). It is a process which a machine authenticates the claimed identity of a person from his or her voice characteristic. Speakers known to the system who claim their true identity are called *claimants*; speakers, either known or unknown to the system, who pose as other speakers are called *impostors* (Bimbot *et al*., 2004).

For text independent speaker verification, Gaussian mixture model (GMM) (Bimbot *et al*., 2004; Reynolds *et al*., 2000; McLachlan and Peel, 2000) and support vector machine (SVM) (Dehak *et al*., 2007) have shown to be the most effective modeling techniques.

In the GMM approach, speaker models are obtained from the adaptation of a universal background model (UBM) through the maximum *a posteriori* (MAP) criterion (Reynolds *et al*., 2000). The UBM is usually trained by means of the expectation-maximization (EM) algorithm from a background dataset, which includes a wide range of speakers, languages, communication channels, recording devices, and environments. The GMM-UBM becomes a baseline technique for text-independent speaker verification due to its reliable performance. A key method in this approach is to use a GMM supervector (Cambell *et al*., 2006) consisting of the stacked means of the mixture components. This GMM supervector can be used as input to SVM to get a new hybrid system namely GMM-SVM.

This paper discusses also the hybrid PCA-GMM-SVM for text independent speaker verification task, where, the new feature vectors with reduced dimension are obtained by applying PCA (Hanilci and Ertas, 2011) to each speaker vector. The PCA is utilized to reduce the dimensions of input vectors firstly for the purpose of reducing the computational complexity, and to lead to more robust estimation of the model parameters.

The reminder of this paper is organized as follows. After reviewing PCA and GMM in Section 2 and Section 3, we discuss the principles of SVM and GMM-SVM in Section 4. Experimental evaluation and results are presented in Section 5 and 6. Finally, conclusions are drawn in Section 7.

2. PCA algorithm

PCA algorithm (Hanilci and Ertas, 2011) also known as Karhunen-Loeve Transform is a linear orthogonal transform method and a powerful technique for feature extraction and dimension reduction. It projects the high dimensional data onto lower dimensional orthogonal space. In the following, we briefly described the PCA algorithm.

Suppose that X is the set of *n* dimensional feature vectors, $x = \{x_1 x_2,...,x_t\}$ the covariance matrix of X is calculated as:

$$C = \frac{1}{T}(X - \overline{X})(X - \overline{X})^T \qquad (1)$$

where $\overline{X}$ is the sample mean of X. Let $\{\lambda_1, \lambda_2,..., \lambda_n\}$ be the eigenvalues of covariance matrix, C, ordered from largest to smallest and $\Phi = \{w_1, w_2,....,w_n\}$ be the corresponding eigenvectors. The matrix Φ is defined as the transformation matrix which projects the original data X into orthogonal feature space. The feature vectors are transformed as (Lee, 2004):

$$Y = \Phi X \qquad (2)$$

The first principal component captures the majority of the variance of the original data, and each successive principal component captures less. The total variance is equal to the original data variance. Dimensionality reduction is mad by keeping some number of the principal components that capture most of the variance in the data set, and discarding the rest [9]. So, the transformation matrix Φ will consist of first L eigenvectors which is associated with largest L eigenvalues, where L is the new dimension (Jason and Manic, 2010).

3. GMM

In the Gaussian Mixture Model (*GMM*) (Reynolds *et al*., 2000), the distribution of the parameterized speech vector of a speaker is modeled by a weighted sum of Gaussian densities:

$$p(x/\lambda) = \sum_{i=1}^{M} p_i b_i(x) \qquad \text{with} \qquad \sum_{i=1}^{M} p_i = 1 \qquad (3)$$

where x is a D-dimensional cepstral vector, λ is the speaker model, $b_i(x)$, $i = 1,..., M$ are the component densities characterized by the mean μ_i and the covariance matrix $\sum_i$. p_i, $i = 1, ..., M$ are the mixture weights. Each component density is a *D*-variate Gaussian Mixture function of the form:

$$b_i(x) = \frac{1}{2\Pi^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)'\left(\Sigma_i\right)^{-1}(x-\mu_i)\right] \qquad (4)$$

For finding model parameters, $\lambda\{\mu_i, \Sigma_i, \prod_i\}$ $i = 1, ..., M$ we use a Maximum Likelihood (ML) estimation method (Reynolds *et al*., 2000). It gives very good results when the training data sets of speaker are large.

The UBM (Universal Background Model) is a large GMM trained to represent the speaker-independent distribution of features (Reynolds *et al*., 2000), which is constructed from a set of background speakers. All the models are diagonal covariance GMMs (Reynolds *et al*., 2000).

The Expectation–Maximization (EM) algorithm is used to find the UBM model parameters (mean, variance and weight) by pooling the data from all the speakers' utterances. The hypothesized speaker specific model is derived by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation (Reynolds *et al*., 2000). For better performance, only the mean vectors are adapted. The UBM model used in this project is a gender-balanced model with 2048-mixtures.

4. SVM and GMM-SVM

Support Vector Machine (SVM), proposed by Vapnik (Vapnik, 2000), is a powerful tool for data classification. Essentially, SVM is a binary classifier to search for the optimal decision boundary in two classes of data, which is based on the principle of structural risk minimization. The new Fisher feature vectors of target speaker and imposters are used to train SVM, so the class decision function for each speaker can be obtained as following:

$$f(x) = class(x) = sign\left[\sum_{t=i}^{N} \alpha_i y_i K(x, x_i) + b\right] \qquad (5)$$

Here $y_i = \{-1, +1\}$ are the ideal output values $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\alpha_i \geq 0$, the support vectors x_i their corresponding weights α_i and the bias term b, are determined from a training set using an optimization process. The target values are either 1 or -1 depending upon whether the corresponding support vector is in class 1 or class 2. For classification, a class decision is based upon whether the value *f(x),* is above or below a threshold. *K(.,.)* is the kernel function. The feature vectors, which are extracted by the proposed method, are nonlinear. So, if they are processed by SVM, the kernel function is needed. In this paper, we selected the Radial Basis Function (RBF) kernel which is more like GMM (Wan and Renals, 2005).

The kernel function allows computing inner products of two vectors in the kernel feature space. In a high-dimensional space, the two classes are easier to separate with a hyperplane. Intuitively, linear hyperplane in the high dimensional kernel feature space corresponds to a nonlinear decision boundary in the original input space (e.g. the MFCC space). For more information about SVM and kernels, refer to (Bishop, 2006).

A GMM supervector is constructed by stacking the means of the adapted mixture components (Cambell *et al*., 2006) from the UBM model. Using the same procedure as that in the GMM–UBM system, each client training supervector is assigned a label of +1 whereas the set of supervectors from a background dataset representing a large number of impostors is given a label of -1. The procedure used for extracting supervectors in the testing phase is exactly the same as that in the training.

5. Experimental protocol

5.1. Description of database

The corpus used in this work is issued from the TIMIT database. This database includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech file for each utterance and is recorded in "ADC" format, where each sentence has 3s of length spoken in English language and divided in 8 dialects.

It consists of a set of 8 sentences spoken by 491 speakers. We have selected 5 phonetically rich sentences for training task and 3 utterances for testing task.

In order to simulate the impostors, a gender balanced UBM consisted from 200 unknown speakers (100 male and 100 female) were trained. The model used 2048 mixture components and was trained using EM algorithm. The full background training dataset was made with five sequences spoken in English by each speaker.

5.2. Parameterization phase

In parameterization phase, we specified the feature space used. This space is defined by vectors of fixed size. Indeed, as the speech signal is dynamic and variable, we presented the observation sequences of various sizes by vectors of fixed size, each vector is given by the concatenation of the Coefficients Mel Cepstrum MFCC (12 coefficients). We also incorporate in the vectors some dynamic information, using the Δ and $\Delta\Delta$ parameters extracted from the middle window every 10 ms. A cepstral mean subtraction (CMS) (Kinnunen and Li, 2010) is applied to these features in goal to fit the data around their average.

In order to calculate the classification function class (x) in SVM model, we used the RBF kernel.

To reduce the dimension of the feature vectors, Principal Component Analysis method performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

5.3. Modeling phase

The GMM-SVM method works by using the GMM supervector consisting of the stacked means vectors of MAP-adapted GMM that captures the acoustics characteristic of a speaker, the supervector is then presented to a speaker-dependent SVM for scoring.

For PCA-GMM-SVM, we applied the PCA dimensionality reduction to the MFCC feature vectors in the front end part.

5.4. Classification phase

In speaker verification there are two kinds of errors; *false acceptance* (FA) and *false rejection* (FR). A FA error indicates when an impostor is classified as an authentic system user. The FR error occurs when an authorized user is wrongly classified as an impostor. There is generally a tradeoff between FA and FR errors.

The point where FR rate is equal to the rate of FA's is known as the Equal Error Rate (ERR) (Martin *et al*., 1997).

6. Experiment results

6.1. Speaker verification using GMM-SVM

To evaluate the influence of dialect and size of database on performance of GMM-SVM, we used TIMIT like corpus of evaluations. The following Table 1 presents the results in term of EER of different dialects and different lengths of subdatabase containing in TIMIT dataset.

Subdataset	Dialect	Number of speakers	EER (%)
dr1	New England	47	14.84
dr2	Northern	90	6.34
dr3	North Midland	86	6.62
dr4	South Midland	65	8.75
dr5	Southern	65	8.45
dr6	New York City	47	14.54
dr7	Western	66	8.37
dr8	Army Brat (moved around)	25	22.07

Table 1: EERs in speaker verification with GMM-SVM using differents subdataset of TIMIT corpora

From Table 1, we noticed that the dialect of speakers has a small influence on EERs in speaker verification, for example, the EER in dr1 (Dialect: *New England*, Number of speaker: *47*) is 7,81% and in dr6 (Dialect is: *New York City*, Number of speaker is: 47) EER=14.54%. However, the number of speakers has a big influence on speaker verification rate. Other wise to say, more the number of speakers is important, more the EER is minor, it's clearly seen with dr8 (25 speaker), the EER= 22.07% where in dr2 (90 speaker), the EER= 6.34%.

6.2. Speaker verification using PCA- GMM-SVM

The main goal of the experiments doing in this Section is to evaluate the verification performances of GMM-SVM using the PCA dimensionality reduction. The results are shown in Table 2.

Subdataset	Dialect	Number of speakers	EER (%)
dr1	New England	47	5.68
dr2	Northern	90	4.16
dr3	North Midland	86	3.58
dr4	South Midland	65	4.18
dr5	Southern	65	4,66
dr6	New York City	47	5.39
dr7	Western	66	4,24
dr8	Army Brat (moved around)	25	8.19

Table 2: EERs in speaker verification with PCA-GMM-SVM using differents subdataset of TIMIT corpora

The main goal of the experiments doing in this Section is to evaluate the verification performance of GMM-SVM using the PCA dimensionality reduction. The results are shown in Table 2.

From table 1 and 2 we can observe that using PCA dimensionality reduction on feature vectors leads to increase the accuracy rate for GMM-SVM approach. It can be explained by the redundancy of the speech signal (Kinnunen and Li, 2010). Otherwise, the speech signal is characterized by elements with high information comparing to other samples which are least significant and are repeated along the signal. But using PCA principal component analysis, it remap the input data into lower dimensional space in such a manner we can eliminate, in the mapped space, those components which are contributed to very low variance, which explain the significant improvement of GMM-SVM performance using PCA.

7. Conclusion

This paper examines the influence of dialect and size of the database on the performance of automatic speaker verification system. It is noticed that the dialect of speakers which speak the same language did not have a big influence on performance of speaker verification task when we used GMM-SVM as a model of training and scoring.

However, we noticed that the size of database affected the performance of verification system. Other wise to say, we obtained a good performance when the size of database (number of speakers) is important than when we have a few speakers in database.

In other hand, we have evaluated the influence of applying PCA dimensionality reduction on the MFCC feature vectors. It has been found out that PCA improves significantly the recognition accuracy.

Références

Reynolds D. A. (2002). An overview of automatic speaker recognition technology. ICASSP, pp. 4072-4075.

Bimbot F. Bonastre J. F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz D., Reynolds D. (2004). A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 4, pp. 430-451.

Reynolds D., Quatieri T., Dunn R. (2000). Speaker verification using adapted gaussian mixture models. Digital Signal Process. 10 (1): 19-41.

McLachlan G. Peel D. (2000). Finite mixture models. Wiley-Interscience.

Dehak R., Dehak N., Kenny P., Dumouchel P. (2007). Linear and non linear kernel GMM supervector machines for speaker verification. Proc. Interspeech, Antwerp, Belgium, pp. 302-305.

Cambell W. M., Sturim D. E., Reynolds D. A., Sololonoff A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. Proc, IEEE Conf.Acoustics, Speech and Signal Processing (ICASSP'06), Toulouse, France, pp. 97-100.

Hanilci C., Ertas F. (2011). VQ-UBM Based Speaker Verification Through Dimension Reduction Using Local PCA. 19th European Signal Processing conference (EUSIPCO'11), Spain.

Lee K.Y. (2004). Local fuzzy PCA based GMM with dimension reduction on speaker identification.

Jason L., Manic W. M. (2010). The Analysis of Dimensionality Reduction Techniques in Cryptographic Object Code Classification.

Vapnik V.N. (2000). The Nature of Statistical Leaning Theory. Springer, New York, second edition.

Wan V., Renals S. (2005). Speaker verification using sequence discriminant support vector machines. IEEE Trans. Speech Audio Process. 13 (2): 203-210.

Bishop C. (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, New York.

Kinnunen T., Li H. (2010). An overview of text independent speaker recognition: From features to supervectors. Speech Communication 52 (1): 12-40, 2010.

Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. (1997). The DET Curve in Assessment of Detection Task Performance. Proc. Eurospeech'97, Rhodes, Greece, September, vol. 4, pp. 1899-1903.

La Transcription Phonétique de la Langue Amazighe : Vers un Système de Synthèse de la Parole*

Hassan Jaa[1] **Belhaj El Graini**[2]
[1] Institut Royal de la Culture Amazighe, BP 2050, Rabat
jaa@ircam.ma
[2] Ecole Mohammadia d'Ingénieur, BP 765, Av Ibn Sina, Rabat
elgraini@emi.ac.ma

Résumé

La synthèse vocale est la passerelle-clé entre le monde de l'écrit et celui de l'oral. Plusieurs domaines d'application s'intéressent au système de synthèse de la parole, l'apprentissage des langues est l'un de ces principaux domaines. Dans cet objectif, ce papier présente la démarche adoptée pour la conception d'un système de transcription phonétique de la langue amazighe qui constituera une des composantes de notre projet qui traite les environnements de traitement automatique des langues pour l'enseignement de la langue et de la linguistique amazighe.

1. Introduction

Dans un contexte socio-économique de compétitivité et de mondialisation, le besoin d'apprentissage continu des langues s'impose comme composante essentielle du système de l'entreprise. Toutefois, les contraintes personnelles et sociales entravent la mise en place de ces processus. Pour répondre aux besoins de formation de la linguistique tout en tenant compte des différentes contraintes, le recours à divers catégories d'outils de traitement automatique des langues a marqué une nouvelle ère d'apprentissage de la langue. L'Institut Royal de la Culture Amazighe (IRCAM) mène des recherches sur le traitement de la langue (analyse et synthèse), pour lesquelles un système de transcription phonétique est un outil indispensable. Notre projet s'inscrit dans le cadre de la mise en place d'un environnement de traitement automatique des langues (TAL) pour l'apprentissage de la langue amazighe.

Ce papier présente la démarche adoptée pour la conception d'un système de transcription phonétique de la langue amazighe. Nous présentons la langue amazighe, sa transcription phonétique et ses problèmes. Nous exposons les composantes d'un système de synthèse vocale et le principe de fonctionnement ainsi que les phases de traitement constituant notre système. Nous allons par la suite décrire l'approche utilisée pour la réalisation de notre application, à savoir, l'approche par règles. Dans cet article, on se limitera à la présentation

* Ce papier représente une contribution présentée au 4ème atelier international sur l'Amazighe et les NTICs sur « Les ressources langagières : construction et exploitation » organisé les 24 et 25 février 2011.

d'une étape du projet, soit la partie concernant la transcription du texte en langue amazighe. Ce système va prendre en considération les variations de prononciation des trois variétés du Maroc : tarifite, tamazighte et tachlhite. Enfin, nous proposerons quelques exemples de règles utilisées par l'application.

2. Graphie et phonétique de l'amazighe

2.1. Graphie de l'amazighe : le tifinaghe

La langue amazighe existe sous forme de dialectes répartis en plusieurs parlers. Avec la création de l'Institut Royal de la Culture Amazighe (IRCAM), la langue amazighe est devenu un standard, et le système graphique de l'amazighe standard proposé par l'IRCAM a permis de neutraliser, sur le plan de l'écrit, certaines réalisations phonétiques non pertinentes entre les trois zones et, au sein d'un même dialecte, entre les différents parlers ; étant entendu qu'une norme graphique ne présuppose nullement l'éradication des variétés régionales. Ce système a une tendance phonologique, en ce sens qu'il ne retient pas toutes les réalisations phonétiques produites, mais uniquement celles qui sont fonctionnelles.

27 consonnes	les labiales (ⵠ, ⵖ, ⴼ) ; les dentales (ⵜ, ⵟ, ⴹ, ⵍ, ⵏ, ⵇ, ⵒ) ; les palatales (ⵛ, ⵊ) ; les alvéolaires (ⵙ, ⵣ, ⵥ, ⵚ) ; les vélaires (ⴽ, ⵅ) ; les labiovélaires (ⴽⵯ, ⵅⵯ) ; les uvulaires (**q, x, v**) ; les pharyngales (ⵄ, ⵃ) et la laryngale (ⵀ) ;
2 semi-consonnes	ⵢ et ⵡ
4 voyelles	trois pleines ⴰ, ⵉ, ⵓ une neutre (ou schwa) ⴻ

Tableau 1 : Le système graphique de l'amazighe

Certaines unités phoniques qui sont soit des variantes régionales, soit des unités non distinctives, soit des unités phonématiques peu productives ne sont pas retenues dans le système.

2.2. Transcription phonétique de la langue amazighe

En amazighe, il n'existe pas de locuteur « standard ». Cette langue se répartit en trois variétés régionales avec tarifite dans le Nord, tamazighte dans le Maroc central et le Sud-Est et tachelhite dans le Sud-Ouest et le Haut-Atlas. Ces dialectes se distinguent par des variations de vocabulaire et une prononciation différente des lettres.

Le tableau suivant représente l'ensemble des caractères de la langue amazighe, tifinaghe, avec pour chaque graphème, sa transcription phonétique standard, ainsi que la transcription phonétique relative à chacune des régions précitées :

Tif	API (Standard)	Tarifit	Tamazight	Tachelhit
ⵄ	a	a	a	a
ⴱ	b	b	b	b
ⴳ	g	ɣ / ʒ	ɣ / ʒ	g / ɣ
ⴳⵯ	g^w	g^w	g^w	g^w
ⴷ	d	d / ð	d / ð	d / z
ⴹ	¨$d^ʕ$	¨$d^ʕ$	¨$d^ʕ$ / $t^ʕ$	¨$d^ʕ$
ⴻ	ə	ə	ə	-
ⴼ	f	f	f	f
ⴽ	k	ʃ /ç/ j	ʃ /ç/ j	k / ç
ⴽⵯ	k^w	k^w	k^w	k^w
ⵀ	h	h	h	h
ⵃ	ħ	ħ	ħ	ħ
ⵄ	ʕ	ʕ	ʕ	ʕ
ⵅ	χ	χ	χ	χ
ⵇ	q	q	q	q
ⵉ	i	i	i	i
ⵊ	ʒ	ʒ	ʒ	ʒ
ⵍ	l	l / r	l	l
ⵎ	m	m	m	m
ⵏ	n	n	n	n
ⵓ	u	u	u	u
ⵔ	r	r /(-)	r	r
ⵕ	$r^ʕ$	$r^ʕ$	$r^ʕ$	$r^ʕ$
ⵖ	ɣ	ɣ / χ	ɣ / χ	ɣ / ħ
ⵙ	s	s	s	s
ⵚ	$s^ʕ$	$s^ʕ$	$s^ʕ$	$s^ʕ$
ⵛ	ʃ	ʃ	ʃ	ʃ
ⵜ	t	θ / t	θ / t	t
ⵟ	$t^ʕ$	$t^ʕ$	$t^ʕ$	$t^ʕ$
ⵡ	w	w /(-)	w /(-)	w
ⵢ	j	j/ ʃ	j/ ʃ	j
ⵣ	z	z	z	z
ⵥ	$z^ʕ$	$z^ʕ$	$z^ʕ$	$z^ʕ$

Tableau 2 : Allures phonétique de l'alphabet Tifinaghe-IRCAM

Une analyse rapide de ce tableau permet de constater qu'à chaque graphème correspond un phonème (transcription standard). Toutefois, la transcription phonétique de certains graphèmes change selon la variante dialectale des trois régions précitées, et parfois on peut avoir différentes prononciations au sein d'un même dialecte. Parmi ces variations :

Dans les régions du Nord et du Maroc central (tarifite et tamazighte), certains caractères peuvent être prononcés de différentes manières, selon leur position dans le mot. Par exemple :

- Le ⵣ qui se prononce /ʒ/ ou /ʒ/ ; exemple : ⴰⵣⵍⴰⵎ (cheval) dont le ⵣ se prononce /ʒ/ : ⴰ/ʒ/ⵍⴰⵎ.
- Le ⵏ qui devient parfois un /ð/
- Le ⵔ qui peut prendre plusieurs formes phonétiques : /ʃ/, /ç/ ou /j/.
- Quelques fois, plusieurs unités phoniques peuvent résulter d'une mutation phonétique, par exemple ⵜⵎⵎⵜ (ma fille) se prononce ⵜ/ʤ/ⵜ (la gémination de ⵎ produit la représentation phonétique ʤ)

Nous allons essayer de prendre en considération ces variations lors de la mise en place du système de transcription, en traitant chaque variante à part, et en concevant une base de données de règles par région ou par variante.

3. Composantes du système de la synthèse vocale

Le système que nous somme entrain de concevoir sera composé de deux modules essentiels : le premier concerne le traitement automatique des langues (TAL), et la seconde intéresse le traitement du signal ou le traitement acoustique. La figure suivante illustre le principe de ce système :

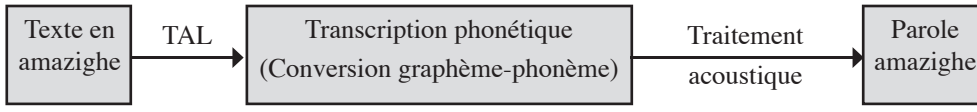


Figure 1 : Les principales composantes du système de synthèse vocale

Le premier module permet de générer une transcription phonétique à partir d'un texte écrit, ce processus de conversion comprend généralement trois étapes essentielles :

- Pré-processing ou prétraitement du texte : son rôle est de diviser le texte en mots, symboles et groupe de mots fortement liés, et de supprimer les caractères parasites (espaces, caractère spéciaux, …), c'est l'étape de la segmentation. Les expressions non alphabétique couramment utilisées sont transformées en mots : les chiffres, les dates, les numéros de téléphone, les e-mails, etc.

- La consultation d'un dictionnaire des exceptions constitué d'un lexique de mots spéciaux.
- L'application des règles de transcription préétablies pour la langue amazighe, et la conversion graphèmes-phonèmes.

La partie qui sera traitée par la suite est le module de traitement linguistique, dont l'objectif est l'aboutissement à un texte phonétique exploitable lors de la phase de traitement acoustique.

4. Principe du système

Ci-après, nous allons décrire les différentes phases de traitement constituant notre système de transcription.

4.1. Segmentation et conversion

Cette phase permet de produire des chaînes de caractères homogènes. En effet, à l'issue de cette étape du traitement, nous aurons en sortie des séquences de même nature : alphabétique, numérique, mots de langues étrangères (généralement le français), et les caractères spéciaux. Les séparateurs utilisés sont le blanc et le retour vers une nouvelle ligne. Les séparateurs : virgules, points, points d'interrogation et d'exclamation etc. sont traités pour analyser les arrêts et les pauses.

L'ensemble des segments issus de la phase précédente sont testés pour savoir s'il faut procéder à une opération de conversion. Dans cette phase, seront transformées en chaînes alphabétiques toutes les expressions non alphabétiques couramment utilisées : les chiffres, les pourcentages, les dates, …

4.2. Dictionnaire d'exception

Ce dictionnaire permettra au système de reconnaître les expressions qui ne peuvent pas être généralement traitées par des règles. En effet, on ne va pas se permettre d'élaborer un nombre important de règles pour traiter quelques mots. Toutefois, si le coût d'exploitation de ce lexique devient exorbitant, on doit, à ce moment là, ressortir des règles spécifiques pour certaines exceptions.

4.3. Approche de traitement

Pour convertir le texte en une transcription phonétique, deux approches sont utilisées généralement : l'approche par règles et l'approche statistiques basée sur l'apprentissage. La nature statistique de la dernière approche nécessite de disposer d'une grande quantité de ressources pour réaliser l'apprentissage (grands corpus de paroles, dictionnaires de prononciation), et ces ressources ne sont cependant pas disponibles directement pour des langues peu dotées comme la langue amazighe. Pour cette raison, nous avons opté pour l'approche par règles, en essayant de mettre en place une base de règles par variante.

4.4. Règles de transcription

Suite à une étude des différents parlers, on a pu définir un certain nombre de catégories, qui vont nous faciliter par la suite l'écriture des règles de transcription. Ces catégories sont :

- L'épellation : elle sera utilisée lors de la prononciation des URL et des Mails (langue étrangère : français) ;
- La prononciation des géminés ;
- Le changement de la prononciation d'un certain nombre de graphème à l'intérieur d'un mot en fonction du contexte dans lequel il est présenté ;
- Les liaisons inter-mots ;
- Le phénomène de l'élision au début et à la fin d'un mot.

Une analyse linguistique des caractéristiques phonétiques des variantes régionales de l'amazighe nous a permis de concevoir quatre (4) familles de règles selon le dialecte : standard, du nord, du centre ou du sud. Cette démarche va, non seulement, faciliter le choix d'une variante, mais aussi rendre le processus d'ajout ou de modification des règles plus simple. La figure suivante illustre le principe adopté :

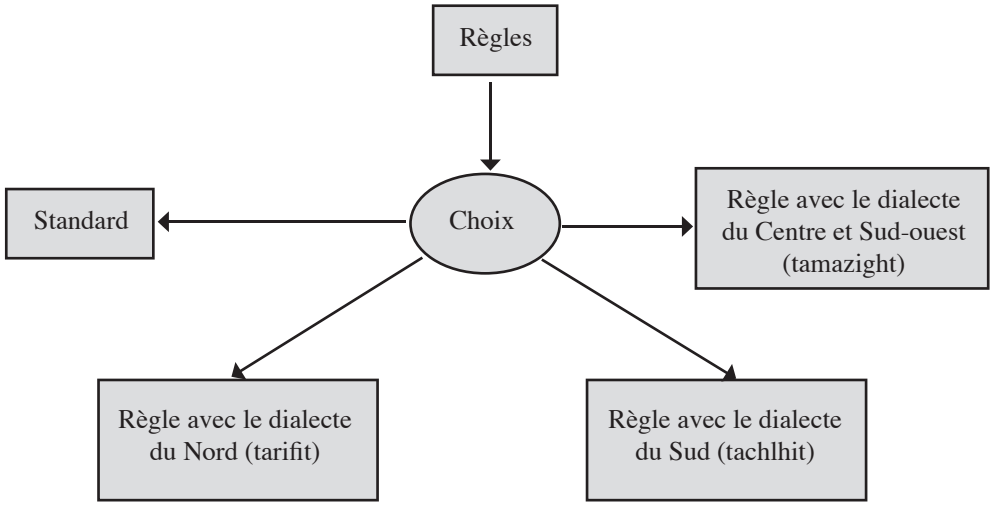


Figure 2 : Le principe du choix du fichier de règles selon la variante

Il faut noter que l'ordre d'application des règles est très important. Donc, on teste d'abord les règles les plus spécifiques avant d'essayer les plus générales.

4.5. Transcription phonétique

C'est la dernière étape du processus de traitement linguistique, elle concerne le passage du texte écrit vers sa transcription phonétique. La transcription utilise le catalogue de règles déjà établies, accompagné de la liste d'exceptions (dictionnaire d'exceptions). Pour chaque graphème sera utilisée sa transcription phonétique adéquate.

5. Présentation de règles

Dans cette partie, nous allons décrire quelques règles utilisées par le système. Pour simplifier la lecture de ces règles, nous utilisons la syntaxe suivante :

règle (n°, "contexte gauche", "graphème", "contexte droit") à ➔ "phonème"

Avec "n°" le numéro de la règle ; C est une consonne, V est une voyelle, L est une lettre quelconque (voyelle ou consonne), EM est un emphatique, $ est le début d'un segment et # est la fin du segment.

D'une manière générale, les règles peuvent indiquer un phonème (un graphème) ou des classes de phonèmes (graphèmes) afin de minimiser le nombre de règles (regroupement par famille). Le numéro de la règle permet de traiter les règles dans un ordre croissant.

- règle (5, "l", "ⵔ", "#") ➔ "/a/"
- règle (6, "l", "ⵔ", "C#") ➔ "/a/"

Lorsque le « ⵔ » est précédé par un graphème quelconque, et qu'il est en fin de mot, on obtient le phonème /a/. Cette règle s'applique uniquement en *tarifite*. De même, lorsque la lettre « ⵔ » est suivie d'une consonne en fin du mot, il devient phonétiquement un /a/. Par exemple : ⴰⵎⴰⵔ et ⵜⴰⵎⵓⵜ se prononcent respectivement ⴰⵎⴰ/a/ et ⵜⴰⵎⵓ/a/t.

- règle (15, "EM", "ⵉ", "l") ➔ "/e/"
- règle (16, "EM", "ⵓ", "l") ➔ "/o/"

Ces deux règles décrivent le comportement des deux voyelles « ⵉ » et « ⵓ » lorsque elles sont précédées pas une emphatique, le graphème ⵉ se transcrit phonétiquement comme un /e/ et le ⵓ devient un /o/.

- règle (24, "a$", "ⵓ ", "l") ➔ "/w/"
- règle (24, "a$", "ⵉ", "l") ➔ "/y/"

Ces deux règles décrivent une réalisation phonétique qui se produit lors du contact des voyelles, c'est la resyllabisation des voyelles hautes. En effet, les voyelles hautes ⵉ et ⵓ se réalisent sur le plan phonétique respectivement /y/ et /w/, dans un contexte vocalique. Par exemple : ⵉⵍⴰ ⵉⵣⵎⴰⵏ devient ⵉⵍⴰ/y/ⵉⵣⵎⴰⵏ et ⵉⴰⴰⴰ ⵓⵔⵅⴰⵣ devient dans le plan phonétique ⵉⴰⴰⴰ/w/ⵔⵅⴰⵣ.

Expérimentation

Pour évaluer notre système de transcription graphème-phonème, nous allons procéder à l'utilisation de plusieurs jeux de données en entrée : article à partir du web (essentiellement le site de l'IRCAM), texte littéraire, et poème. Il sera testé aussi pour vérifier la prise en charge des trois variantes (selon le choix du fichier de règles). La mesure d'évaluation qu'on va utiliser sera la précision en mot : nombre de mots parfaitement et entièrement phonétisés sur le nombre total de mots dans le texte d'entrée. Ces vérifications vont nous permettre d'améliorer le système en précisant l'importance de chaque règle.

6. Conclusion et perspectives

Dans cet article, nous avons présenté les éléments de base ainsi que les approches utilisées pour la mise en place d'un système de transcription phonétique de la langue amazighe. Ce système présente une certaine flexibilité puisque les règles sont stockées sur des fichiers à part, ce qui facilite le changement de la variante d'une part, et la révision des règles d'autre part, sans une recompilation du programme.

Cependant, les fichiers des règles de transcription doivent s'enrichir en vue d'obtenir de meilleurs résultats, une tâche qui nécessite une analyse plus approfondie de la phonétique de l'amazighe.

Un cas que nous n'avons pas abordé au cours de ce travail est celui du "`I`" emphatique, car cette unité permet la production d'une opposition phonémique : `II8` sans emphase signifie sentir bon, `II8` avec emphase signifie sentir mauvais. C'est une unité peu productive et très localisée, et son traitement linguistique nécessite une analyse sémantique.

Références

Ameur M., Bouhjar A., Boukhris F., Boumalek A., El Medlaoui M., Iaazzi E., Souifi H. (2004). *Initiation à la langue amazighe*, Rabat, Publications de l'IRCAM.

Ansar, K. (2005). Sibilants in *Berber*, Doctoral dissertation, Mohammed V University, Rabat.

Béchet F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement Automatique des Langues - TAL*, vol. 42, pp. 47-67.

Black A. W., LENZO K., PAGEL V. (1998). Issues in building general letter to sound rules. *In Proceedings of the 3rd ESCA Workshop in Speech Synthesis*, Jenolan Caves, Australie.

Boukhris F., Boumalek A., El Moujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*, Rabat, Publications de l'IRCAM.

Boukous A. (2009), *Phonologie de l'amazighe*, Rabat, Publications de l'IRCAM.

Lemmety S. (2000), Review of speech synthesis technology, *Helsinki University of Technology*. Thèse.

Moudenc T., Emerard F. (2003). Synthèse vocale et handicap. *Annales de télécommunications*. pp. 928-934.

Moulines E., Cappe O. (1996). Synthèse de la parole à partir du texte. *Techniques de l'ingénieur*. H1960, pp. 7.

TICAM'12, p. 173-182

Reformulation des Requêtes pour le Domaine de l'Enseignement, Cas des Cours de l'Algorithmique

Mohamed Rachdi El Habib Ben Lahmar El Hassan Labriji
Aziz Chara Kamel Et Guemmat
Faculté des sciences Ben M'sik, Casablanca, Maroc
{mohamed.rachdi, labriji}@yahoo.fr
{h.benlahmer, k.elguemmat}@gmail.com
aziz.chara@hotmail.fr

Résumé

La croissance très importante des informations disponibles sur Internet nécessite la mise en place des outils de recherche de plus en plus performants, permettant de retrouver des documents pertinents parmi un ensemble de documents disponibles. La sélection des documents adéquats à une requête, nécessite forcement le passage par plusieurs étapes. Parmi ces étapes, on trouve l'étape de la reformulation. C'est un processus qui permet d'enrichir la requête utilisateur afin de borner le résultat souhaité et donc, de guider l'utilisateur vers le bon document. Dans la littérature, on trouve plusieurs techniques de reformulation, notamment la reformulation par réinjection.

Le propos de cet article entre dans le même cadre, son objectif est de proposer une nouvelle approche d'enrichissement des requêtes, en se basant sur des classes bien définies. L'enrichissement qu'on propose sera appliqué, dans le système de requêtage d'un méta moteur de recherche dédié au domaine de l'enseignement. Nous détaillerons en particulier le cas du cours d'algorithmique.

Notre approche, définira un ensemble de classes distinguées en deux catégories : les classes d'application qui représente des concepts du cours d'algorithmique et les classes d'utilisation constituant des patrons liés au classes d'application et regroupant chacune une collection des marqueurs (métalinguistiques, métalinguistiques nominaux, lexicaux, …). Ces classes servent à détecter le besoin de l'utilisateur dans un premier temps, et permettant aussi de reformuler la requête utilisateur et de l'enrichir. Les différentes classes sont complémentaires et leur utilisation permet d'augmenter la pertinence des résultats retournés.

1. Introduction

La reformulation des requêtes constitue dans nos jours une des clés de la réussite des Systèmes de Recherche d'Information, car elle a un impacte direct sur le résultat retourné. En effet la qualité du résultat dépend toujours du besoin exprimé.

Un besoin bien reformulé, retourne un résultat bien pertinent avec plus de précision et moins de bruit. Dans le même souci, et parmi les techniques de reformulation existantes, il y a des approches d'enrichissement qui se basent sur, les ontologies de domaine personnalisées (Aimé *et al*., 2010), les profiles utilisateurs (Koutrika et Ioannidis, 2004 ; Koutrika et Ioannidis, 2005).

En effet, (Baeza-Yates et Ribeiro-Neto, 1999) ont défini trois types d'outils de reformulation de requêtes qui dépendent des données utilisées : des données provenant du jugement de la pertinence par l'utilisateur, des données récupérées à partir de l'ensemble de documents initialement retournés par le SRI et des données globales provenant d'une collection de documents et/ou de ressources externes.

L'enrichissement présenté dans cet article appartient au troisième type défini par (Baeza-Yates et Ribeiro-Neto, 1999) et a comme objectif : augmenter la précision et diminuer le bruit. Cet enrichissement qu'on propose s'applique pour un domaine bien défini : l'enseignement, et exactement pour le module algorithmique. Dans notre approche, on se base pour enrichir la requête sur des classes d'application et des classes d'utilisation.

Dans ce qui suit, on présente au début un état de l'art sur les différentes techniques de reformulation de requêtes. Une fois achevée, on passera à la présentation de notre contribution. La partie suivante sera consacrée à l'évaluation de notre système et on terminera par une conclusion et des perspectives.

2. Etat de l'art

La quantité énorme d'information disponible sur Internet et le nombre important des utilisateurs pausent de vrais problèmes pour la sélection des documents pertinents. La reformulation des requêtes constitue donc la solution adéquate pour satisfaire le client (utilisateur final).

Dans le même sens, plusieurs pistes ont été explorées, ainsi l'utilisation des ontologies de domaine représente une parmi les meilleures solutions. Plusieurs travaux existent déjà exploitant les ontologies de domaine pour reformuler les requêtes (Aimé *et al*., 2010 ; Sy *et al*., 2012 ; Soussi *et al*., 2008). Ces approches utilisant leurs propres ressources. D'autres travaux de reformulation exploitant des méthodes linguistiques et des ressources lexicales sont aussi existantes (Med El Amine, 2009), l'apport de ces approches est qu'ils font une analyse morphologique des mots de la requête avant la reformulation. La réinjection par pertinence semble dans ce stade une solution aussi assez répondue, dans le sens ou elle intègre l'utilisateur dans le processus de la reformulation. Des travaux lancés dans le même cadre se basant, pour reformuler la requête, sur des documents jugés pertinents par l'utilisateur (Boughanem *et al*., 1999; Ruthven et Lalmas, 2003), d'autres travaux comme celui de (El Younoussi *et al*., 2010; Ben Lahmar *et al*., 2010) utilise la réinjection par pertinence mais sans intervention de l'utilisateur (automatique), elle sera appliquée à N résultats, le résultat final retourné sera considéré définitif.

Dans (Rachdi *et al*., 2011), une autre technique de reformulation a été proposée, elle se base sur les phrases de définition issues d'Internet. Cette technique a l'inconvénient qu'elle s'applique sur des requêtes composées d'un seul mot.

3. Approche

Actuellement, le souci majeur des systèmes de recherche d'information est la sélection des documents pertinents parmi une masse énorme de documents existants sur Internet. Donc il parait que la tâche n'est pas facile. Pour l'accomplir correctement, plusieurs techniques de reformulation existent déjà, dont on a cité quelques une dans la section précédente. A notre connaissance, jusqu'à maintenant, il n'y a pas des approches de reformulation de requêtes dédiées au domaine de l'enseignement et spécialement pour les cours d'algorithmique.

L'enrichissement qu'on propose dans ce papier se situe au cœur de ces approches de reformulation, et permet de reformuler la requête utilisateur en se basant sur des classes d'application et des classes d'utilisation et des marqueurs définis pour les classes d'utilisation. On a exploité directement l'Internet pour la reformulation.

Considérons donc les définitions suivantes :

Classe d'application : elle représente un élément indispensable du cours d'algorithmique, comme exemple des classes d'application on peut citer: variable, lecture et écriture, structures conditionnelles, structures répétitives, tableaux, fonctions et procédure, ….

Classe d'utilisation : Elle correspond à un patron correspondant à un élément du cours, elle est définie par un ensemble de marqueurs permettant de l'identifier. Comme exemple des classes d'application on peut citer: Définition, Objectifs, déclaration, Affectation (utilisation), syntaxe, types, exemples, ...

Un patron lexico-syntaxique (marqueurs): décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte répondant à ce format. Dans le cas particulier de la recherche de relations, le patron caractérise un ensemble de formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes (Rebeyrolle et Tanguy, 2000).

Comme exemple des patrons, pour la classe Déclaration on peut trouver les marqueurs suivants : syntaxe de déclaration, déclarer, utilisation, déclaration, ….

Notre méthode ramène le problème de la reformulation à la recherche d'un ensemble de mots ayant la même sémantique que les mots de la requête initiale.

Cette approche est sensée fournir l'ensemble des mots aidant à la reformulation. Cependant, l'objectif réel de l'approche est le renvoi du maximum des documents pertinents pour l'utilisateur.

Soit donc un besoin de l'utilisateur exprimé sous forme de requête utilisateur. Comment détecter ce besoin ? Comment faire la correspondance entre ce besoin et les classes définies? Et comment exploiter ces classes pour l'enrichissement ?

Ce papier est donc a pour objectif de répondre à ces questions.

Dans un contexte général, les classes d'application et d'utilisation représentent des éléments prédéfinis qui aident à détecter le besoin de l'utilisateur dans un premier temps et après pour l'enrichissement.

Notre approche consiste donc à détecter les besoins généraux et spécifiques pour enrichir la requête utilisateur.

Le besoin général correspond à l'utilisation des classes d'application qui servent à détecter les éléments du cours d'algorithmique cherché par l'utilisateur (comme exemple des besoins généraux, variables, fonctions, ….

Une fois les besoins généraux détectés, l'ensemble des classes d'utilisation nous aide dans ce niveau à extraire les besoins spécifiques (comme exemple des besoins spécifiques : déclaration, syntaxe, ...).

Dans le cas où les besoins spécifiques sont exprimés implicitement, on le détecte à l'aide des marqueurs des classes d'utilisation.

Notre approche après détection des besoins généraux et spécifiques (classes d'application et d'utilisation) enrichie la requête avec les classes d'application trouvées et les marqueurs liés aux classes d'utilisation trouvées.

Le schéma suivant montre le processus général de l'approche.

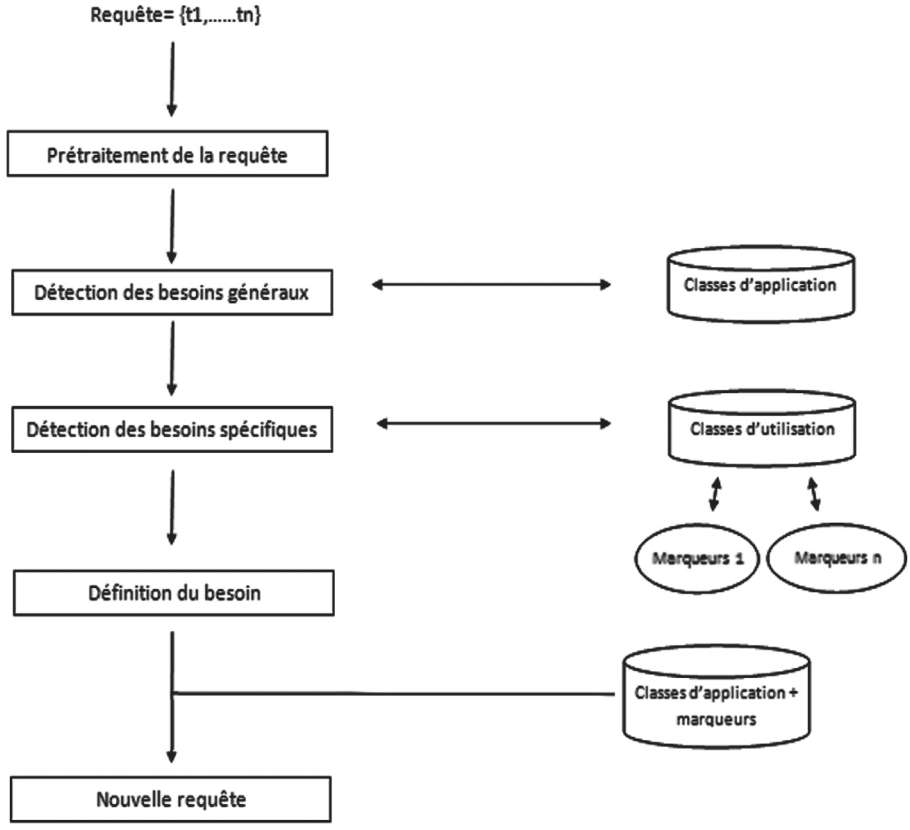


Figure 1 : Processus de l'approche de l'enrichissement par classes

Pour mettre en place ce processus, on passe par quatre étapes :

3.1. Étape 1 : prétraitement de la requête

Une requête utilisateur est composée d'un ou plusieurs mots, représentés soit séparément soit sous forme d'une phrase, question, …. Afin de comprendre ce besoin on passe par plusieurs étapes dont une étape de prétraitement est primordiale avant de commencer le processus. Ce prétraitement permet de nettoyer la requête des mots vides et ne garder que les mots importants porteurs de sens dans la requête. Pour sélectionner les termes importants représentant la même sémantique que la requête l'élimination des mots vides est donc la première étape du processus.

Soit donc Q_0 : la requête utilisateur.

$$Q_0 = \{t_1, \ldots, t_n\} \qquad (1)$$

avec t_n est le terme n de la requête.

Le prétraitement de la requête nous donne une nouvelle requête Q_1

$$Q_1 = \{ M_1, \ldots, M_n\} \qquad (2)$$

dont M_n est le mot n porteur de sens.

3.2. Étape 2 : détection du besoin général

La première étape nous permet (à l'aide du prétraitement) de dégager les termes importants contenus dans la requête et qui représentent sémantiquement le besoin de l'utilisateur.

Cette deuxième étape est pour objectif de détecter les besoins généraux exprimés dans la requête utilisateur.

On dispose au préalable d'un ensemble de classes d'application et d'utilisation (ECA). Le propos de cette étape consiste à faire une projection des termes de la requête initiale Q_1 sur la liste ECA. Cette projection nous permet de savoir quelle est la classe d'application (CA) qui correspond au plus à la requête Q_1 et spécialement savoir l'élément du cours d'algorithmique cherché par l'utilisateur.

Soit donc ECA, l'ensemble des classes d'application

$$ECA = \{CA_1, CA_2, \ldots, CA_n\} \qquad (3)$$

avec CA_i est la classe d'application i.

L'interrogation de la liste ECA dans ce stade permet de savoir donc les besoins généraux de la requête. Le résultat de cette étape est l'ensemble des classes d'applications adéquates à la requête Q_1.

La prochaine étape sera donc la détection des besoins spécifiques de l'utilisateur afin d'affiner la phase de détection des besoins.

3.3. Étape 3 : détection du besoin spécifique

Le besoin de l'utilisateur n'est pas toujours bien exprimé et souvent renseigné implicitement. Dans ce sens, pour savoir exactement ce besoin, on a commencé par le prétraitement de la requête et on a achevé par la détection du besoin général. La présente étape viendra pour compléter le travail précédent on détectant le besoin spécifique.

Deux cas se présentent dans cette étape :

- Soit le besoin est exprimé explicitement par l'utilisateur alors on peut le détecter directement.
- Soit il n'est pas défini explicitement et donc on fait référence aux marqueurs définis dans les classes d'utilisation.

On commence cette étape par l'interrogation de l'ensemble des classes d'utilisation (ECU) déjà défini. Une fois interrogée, on fait un appariement entre la liste ECU et l'ensemble des termes de la requête Q_1.

Considérons donc, ECU, l'ensemble des classes d'utilisation :

$$ECU= \{CU_1, CU_2, CU_3, \ldots, CU_n\} \quad (4)$$

Soit CU, une classe d'utilisation. Chaque CU est défini par une liste des marqueurs :

$$CU= \{MQ_1, \ldots, MQ_n\} \quad (5)$$

avec MQ_i est le marqueur i défini.

Dans le cas où l'appariement retourne une classe CU, on déduit que cette classe CU représente vraiment le besoin spécifique de la requête.

Dans le cas échéant où l'appariement retourne un résultat nul, on passe à un deuxième niveau d'appariement cette fois ci entre les mots de Q_1 et tous les marqueurs définis dans les classes d'utilisation. Le résultat dans ce dernier cas est un marqueur utilisé dans la requête.

A partir de 5, on a considéré qu'une classe CU est défini par un ensemble de marqueurs, on appliquant cette règle inversement pour déduire la classe CU qui contient le marqueur retourné. Dans ce dernier cas, la classe CU contenant ces marqueurs est considérée donc le besoin spécifique de l'utilisation. A la fin de cette étape, on dispose du besoin général, du besoin spécifique, la prochaine étape sera donc l'exploitation des besoins détectés pour l'enrichissement de la requête.

3.4. Étape 4 : enrichissement de la requête

Dans les étapes précédentes, on a pu détecter le besoin général et spécifique de l'utilisateur. Ces besoins on les exploite dans la présente étape pour enrichir la requête utilisateur.

Dans la règle (5), on a défini une classe CU avec une liste des marqueurs qui lui sont associés. A partir du besoin spécifique (Classe d'utilisation), on extrait tous les marqueurs liés à cette classe. Ces marqueurs on les utilise pour enrichir la requête finale.

La requête finale donc sera constituée du besoin général détecté (étape 2) et de l'ensemble des marqueurs extrait à partir du besoin spécifique (étape 3). La requête finale sera donc :

$$Req = (CA\ MQ_1\ CA\ MQ_2, \ldots, CA\ MQ_n\} \quad (6)$$

Exemple : si l'utilisateur tape « comment déclarer une variable »

1. dans un premier temps, le prétraitement sert à éliminer les mots vides « comment » et « une », il nous reste que les deux termes « variables » et « déclarer ».
2. Dans la phase de la détection du besoin général correspondant à la classe d'application, on détecte que l'utilisateur veut s'informer sur les variables.
3. La détection du besoin spécifique correspondant aux classes d'utilisation permet de déterminer le besoin précis de l'utilisateur, il s'agit dans ce cas du mot « déclarer ».
4. La requête sera enrichie donc avec la classe d'application Variable et les marqueurs de la classe d'utilisation Déclaration. Elle devienne:

req={Déclarer Variable, syntaxe de déclaration de variable, utiliser variable, utilisation des variables, ...}.

Algorithme de l'approche

L'algorithme suivant montre les étapes du processus d'enrichissement :

Entrées: $Q_0 = \{t_1, \ldots, t_n\}$

Sorties: Q'

Req : la requête utilisateur

C_u : tableau des classes d'utilisation

Mq : tableau des marqueurs

On commence par le prétraitement à l'aide de la fonction prétraitement. Ensuite, on réalise l'appariement du premier niveau, entre les mots de la requête et la liste des classes d'application (ECA) pour détecter les besoins généraux. Puis, on passe au deuxième appariement entre les mots de la requête et la liste des classes d'utilisation (ECU) pour extraire les besoins spécifiques. Une fois le besoin utilisateur détecté on enrichie la requête avec les classes d'application et les marqueurs des classes d'utilisation trouvées.

Req=prétraitement(Q_0)

```
Pour  i=0 à |req|
        Si t(i) Є ECA alors
                bgénéral.add(t(i))
        Sinon
                Si t(i) Є ECU alors
                        Pour m=0 à |t(i)|
                        Bspécifique. Add(t(i).mq(m))
                        Fin pour
```

```
                Sinon
                        Pour k =0 à |ECU|
                                Pour j=0 à |Cu(k)|
                                        si t(i) Є mq(j)
                                                pour m=0 à |t(i)|
                                                        bspécifique. Add(t(i)mq(m))
                                                fin pour
                                        fin si
                                fin pour
                        fin pour
                fin si
        fin si
Fin pour
```

Q'={bgenéral+bspécifique}

4. Résultat

Pour tester notre approche, on a utilisé le moteur de recherche Google en tant que plate forme pour la mise en place de notre système. On a utilisé l'exemple cité précédemment d'un utilisateur cherchant comment déclarer une variable en algorithmique. Le système a envoyé comme requête « déclaration variables, déclarer variable, syntaxe déclaration variable, utilisation variables… ». Comme résultat on a obtenue une liste de documents contenants la réponse à la requête « comment déclarer une variable ».

Le test qu'on a fait pour cet exemple a été effectué au début sur 20 premières pages retournées par Google et a montré que sans reformulation on a obtenue 11 documents non pertinents, par contre avec notre approche on a obtenue, parmi les 20 retournés, 7 documents non pertinents.

On a fais le même test avec 100 liens, 47 documents non pertinents sont retournés sans reformulation. Alors qu'avec notre approche de reformulation seulement 26 documents non pertinents sont retournés.

Le résultat retourné avec notre approche semble donc adéquat pour la requête et représente une valeur ajoutée par rapport à l'envoi de la requête sans reformulation.

Exemple de liens pertinents retournés que par notre approche et non retournés sans reformulation :

http://www.charlie-soft.com/Programmation/Algorithmique/index.php

http://mescal.imag.fr/membres/arnaud.legrand/teaching/2003/Algo-DEUG-03/Cours/node3.html

http://www.mathmaurer.com/visiteurs/2nde/algo_01_bases/www.mathmaurer.com-2nde_algo_cours_01_Bases_en_algorithmique.pdf

http://www.fsr.ac.ma/cours/informatique/elmarraki/Algo_ch1_3.pdf

5. Conclusion

Ce travail entre dans le cadre du développement d'un système qui aide à enrichir la requête utilisateur cherchant dans le domaine de l'enseignement et exactement un cours d'algorithmique. Nous avons utilisés la notion de classes d'application, les classes d'utilisation et les marqueurs pour la reformulation.

Le but de ce traitement est d'augmenter la précision et diminuer au maximum le bruit du résultat retourné. La faisabilité de la méthode a été testée en utilisant le moteur de recherche Google.

Nous avons montré que notre approche représente une valeur ajoutée dans la phase de requêtage du processus de recherche des systèmes de recherche d'information.

Malgré que notre approche est prometteuse, on estime dans un premier temps mettre en place une plate forme pour évaluer notre système en utilisant un corpus, puis une amélioration de l'approche afin de prendre en considération la reformulation des cours d'informatique en générale et pas seulement le cours d'algorithmique. L'utilisation de ces classes pour la construction d'une ontologie de domaine constitue aussi une perspective envisagée après ce travail.

Références

Sy M. F., Ranwez S., Montmain J., Ranwez V. (2012). Une méthode de reformulation utilisant une ontologie de domaine. CORIA 2012. Bordeaux, 21-23 mars 2012, pp. 135–150.

Rachdi M., Ben Lahmer H., Labriji H. (2011). Semantic enrichment of queries in search engines. Extraction et Gestion des Connaissances (EGC-M 2011), novembre 2011, Tanger (Maroc).

Aimé X., Fürst F., Kuntz P., Trichet F. (2010). Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées. Actes de l'atelier de la Personnalisation du Web, 10ème journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010), 2010.

El younoussi Y., Doukkali Sdigui A., Ben Lahmer E. (2010). Promoting the relevance of the research information systems via a query reformulation process. *The 6th International Computing Conference in Arabic (ICCA'10), mai 2010*, Tunes, Tunisie.

Ben Lahmer E., Doukkali Sdigui A., El younoussi Y. (2010). The research of terms definitions by metasearch. *The 6th International Computing Conference in Arabic (ICCA'10), mai 2010*, Tunes, Tunisie.

Med El Amine A. (2009). Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. *The 2nd Conférence Internationale sur l'Informatique et ses Applications (CIIA'09), Saida, Algerie, m*ai 3-4, 2009.

Soussi R., Ben Mustapha N., Zghal H. B., Aufaure-Portier M. (2008). Un système d'aide à la recherche d'information en ligne basé sur les ontologies (SA-RI-Onto). *Conférence en Recherche d'Infomations et Applications, Renne, France*, pp. 483-490.

Koutrika G., Ioannidis Y. E. (2005). Personalized Queries under a Generalized Preference Model. *The 21st International Conference on Data Engineering, Tokyo, Japan*, April 5-8, 2005.

Koutrika G., Ioannidis Y. E. (2004). Personalization of Queries in Database Systems, *The 20th International Conference on Data Engineering, Boston, Massachusetts, USA*, April, 2004.

Ruthven I., Lalmas M. (2003). A survey on the use of relevance feedback for information access systems. Knowl. Eng. Rev., 18(2):95-145, 2003.

Rebeyrolle J., Tanguy L. (2000). *Repérage automatique de structures linguistiques en corpus: Le cas des énoncés définitoires*. Cahiers de Grammaire, N° 25. pp. 153-174

Baeza-Yates R. A., Ribeiro-Neto B. (1999). Modern Information Retrieval, *Addison-Wesley Longman Publishing Co*., Inc., Boston, MA, USA.

Boughanem M., Chrisment C., Soule-Dupuy C. (1999). Query modification based on relevance backpropagation in adhoc environment. Information Processing and Management, N°35. pp. 121-139.

TICAM'12, p. 183-191

Recherche par Expression dans le Texte Intégral Multilingue

Yahya Hlal
Docteur Es-sciences, Ex-Professeur à l'Ecole Mohammadia d'Ingénieurs à Rabat
yh5099@hotmail.com

Résumé

Cet article relève de l'Ingénierie des langues, spécialement celle relative à la langue arabe. Plus particulièrement, cela s'inscrit dans le cadre de la recherche de l'information véhiculée par le texte naturel. La difficulté première dans ce domaine est la non reconnaissance des informations recherchées (notion de silence). On fait état dans ce papier des difficultés susceptibles d'être rencontrées et on propose des solutions pour les résoudre. Ces difficultés vont de la non reconnaissance du caractère, des mots, des concepts et plus généralement des schémas d'articulation entre éléments dans le discours. Les solutions proposées s'appuient sur la notion d'affinités entre les objets. Ces derniers peuvent être élémentaires (caractères) ou plus ou moins complexes comme les mots (niveau morphologique), les concepts (niveau sémantique) ou les schémas d'agencement de ces éléments dans ce que nous avons appelé « schème syntaxique » qui est un concept que nous présentons pour la première fois dans cet article.

Les applications des outils issus de ces concepts sont immédiates et trouvent leur place de façons évidente dans l'exploitation des bases de données textuelles multilingues. Nous les utilisons d'ores et déjà dans le produit GED-IMMA qui est utilisé au Maroc. Mais cela ouvre la perspective d'exploration des connaissances véhiculées par les textes multilingues en vue de les exploiter dans des systèmes du type « Intelligence Naturelle ».

1. Introduction

Nous présentons dans cet article des fonctionnalités de recherche dans le texte intégral multilingue (Arabe et français). Ce qui est attendu dans la recherche dans les corpus de textes est la l'identification des documents dont le contenu est en relation avec des critères de recherche utilisés.

La technique classique consiste à reconnaître les arguments de recherche (mots ou partie de mots) dans le texte. Cela se fait par la reconnaissance explicite des configurations de caractères. Cette technique est simple; mais elle conduit à du silence, aussi intolérable qu'injustifié. En effet,

- on considère deux caractères comme étant différents dès lors que leur code de représentation est différent. Ainsi, le "é" est aussi différent du "è" que du "z". Or, de façon intuitive, on admet qu'il existe une affinité (degré de ressemblance) entre le "é" et "è" laquelle affinité n'existe pas avec le "z".

- On considère deux mots comme étant différents dès lors qu'ils sont constitués de caractères différents. Ainsi, le mot "imposition" est aussi différent du mot "imposable" que du mot "science". Or, de façon intuitive, on admet qu'il existe une affinité (degré de ressemblance) entre le mot "imposition" et le mot "imposable" laquelle affinité n'existe pas avec le mot "science". Il s'agit là d'une affinité d'ordre morphologique qu'il y a lieu d'exploiter pour réduire le silence.
- De la même façon que précédemment, on considère à priori que le mot "impôt" est différent du mot "redevance". Or, de façon intuitive, on admet qu'il existe une affinité (degré de ressemblance) entre ces deux mots. Il s'agit là d'une affinité d'ordre sémantique qu'il y a lieu d'exploiter pour réduire le silence. Cela conduit à l'approche de la recherche par concept qui ouvre la voix vers la recherche multilingue où l'affinité sera étendue à des mots extra-langue ("impôt" = "ضريبة").

On fera état dans cet article des options de recherche qui traitent ces types d'affinités pour doter les moteurs de recherche des possibilités de recherche exhaustive (lutte contre le silence).

Les options de recherche qui peuvent être utilisées de façon autonome, pourront être utilisées dans la recherche par expression qui constitue le point fort de cet article.

2. Fonctionnalités pratiques pour la recherche dans le texte intégral

Les difficultés à résoudre, lorsqu'on se fixe comme objectif le maximum d'exhaustivité (lutte contre le silence), est la reconnaissance des éléments ayant une affinité avec les arguments de recherche (degré de ressemblance).

Par exemple, comment établir les affinités suivantes :

Dans le cas de l'arabe :

تشبث = تشبت
ضريبة = ضرائب
تعليم = تربية

Dans le cas du français :

Impôt = impot
Information = Informationnel
Education = Enseignement

En extra-langue :

Education = Enseignement = تربية = تعليم

Ces difficultés peuvent se répartir selon les axes suivants :

- ***Affinité entre caractères***

Si la différence entre deux lettres comme "A" et "Z" est manifeste; il n'en est pas de même pour les lettres "o" et "ô" où l'affinité est évidente. Ce type de difficulté est résolu par l'emploi du concept "Classe de caractères" qui consiste à confondre les caractères d'une même classe.

Par exemple, on peut établir les classes suivantes :

En arabe :

{د, ذ } ; {ض, ظ} ; {ت, ث}; etc.

En français :

{e, é, è, ê, ... } ; {o, ô} ; {i, ï}; etc.

- ***Affinité morphologique entre mots***

Il est utile de pouvoir établir une affinité entre des mots sur un plan purement morphologique (et non lexical). Par l'exemple, la recherche avec un mot tel que "imposition" devrait permettre de retrouver les textes contenant des mots tels que "imposable", "imposer", etc. On remarque que tous ces mots possèdent la même base "impos". Un traitement au niveau des suffixes permet d'établir une affinité entre la famille de mots "imposition, imposable, imposer, etc.". Ce raisonnement relatif aux suffixes peut être étendu aux préfixes.

- ***Affinité morphologique entre mots arabes***

Le système morphologique arabe basé sur l'emploi d'une racine et d'un schème pour générer les mots permet d'établir de façon évidente une affinité morphologique entre les mots.

Par exemple, la racine "كتب" peut représenter la classe des mots ayant tous cette racine dans leur composition :

كتب ← (كاتب، مكتوب، كتاب، مكتبة)

Un traitement au niveau des racines permet d'établir une affinité entre la famille de mots de mêmes racines.

Il est à remarquer que cette approche permet de résoudre le problème du pluriel brisé (dû à l'emploi d'un infixe) qui devient un cas particulier d'affinité entre mots. C'est ainsi que l'on pourra confondre dans une recherche la forme pluriel et son singulier, via la classe de mots associés à la racine, comme le montre l'exemple suivant :

درس ← (مدرسة، مدارس، دارس، مدروس)

- ***Affinité sémantique (conceptuelle)***

Dans le cadre d'un contexte métier, il est possible d'établir des affinités lexicales (associations explicites établies dans un lexique) entre des mots; de sorte que la recherche par l'un des mots permet de retrouver tous les textes contenant les autres mots. Ainsi, si on établit les associations suivantes :

Machine → machin, engin, appareil

→ آلة، آلات، مكنة، مكنات، جهاز،أجهزة

La recherche avec le mot "engin" permet de retrouver tous les textes qui contiennent :

en intra-langue (français) les mots "machine" , "appareil" , "engin"
en extra-langue (arabe) les mots "آلة" , " مكنة", "جهاز" , "أجهزة", etc.

3. Recherche par expressions

La recherche par expression est une approche qui s'inscrit dans le cadre des recherches contextuelles positionnelles qui relèvent globalement du traitement du type syntaxique. Cette approche de recherche est très riche, vue qu'elle utilise différents aspects aussi bien morphologique, syntaxique que sémantique. Son intérêt est également de permettre de s'intéresser à des classes d'expressions, par l'emploi de moules syntaxiques, où l'on pourra spécifier de façon précise les éléments constants et les éléments variables se trouvant dans un contexte positionnel rapproché (la distance maximale pouvant séparer les éléments est fixée mais paramétrable). Les éléments fixes du moule pourront être :

- Une partie de mot qui peut être contenu dans d'autres mots, tel que "عمل" qui peut se trouver dans "وعمل" , "ويعمل", "ويعملون" , etc.
- Un mot entier qui ne doit pas être contenu dans un autre;
- Une racine qui représente les mots qui peuvent en dériver. Si on spécifie la racine "كتب" alors elle sera remplacé par les mots du type : "كاتب". "كتاب". "مكتوب". "مكتبة". etc.;
- Un mot qui représente la famille des mots ayant une affinité morphologique (emploi des préfixes et suffixes);
- Un mot qui représente la classe des mots ayant une affinité sémantique (emploi d'un lexique de champs sémantiques métiers);

Ces moules syntaxiques; avec ses éléments constants et variables peuvent être assimilés à la notion de schèmes sauf que ces derniers se situent au niveau des mots alors les éléments syntaxiques se situent au niveau d'une partie de discours.

4. Conventions relatives aux critères de recherche par expression

Dans le cadre de la recherche par expression, le système utilise un certain nombre d'options qui constituent le contexte d'exécution des requêtes et qui fixe la signification des éléments utilisés dans les critères de recherche. Nous allons faire état de ces options dans ce qui suit :

4.1. Les options de recherches

- Aucune option particulière : on recherche les mots utilisés comme critères en tant que mots entiers ou contenus dans d'autres mots;
- Recherche de mots entiers : on recherche les mots utilisés comme critères en tant que mots entiers;
- Racines arabes : On fait la recherche avec les mots générés à partir des racines spécifiées comme critère de recherche;

- Classe de caractères : On fait la recherche avec les mots générés sur la base de l'affinité type caractère;
- Préfixe (Famille de mots) : on fait la recherche avec le reste du mot après troncature du préfixe dans le mot (affinité morphologique : emploi du préfixe);
- Suffixe (Famille de mots) : on fait la recherche avec le reste du mot après troncature du suffixe dans le mot (affinité morphologique : emploi du suffixe);
- Concepts arabes : on fait la recherche avec les mots générés sur la base de l'affinité sémantique dans la langue arabe;
- Concepts français : on fait la recherche avec les mots générés sur la base de l'affinité sémantique dans la langue française.

4.2. Distance maximale séparant les mots de l'expression

Elle est exprimée en nombre de mots séparant deux éléments recherchés dans le texte. C'est sur la base de cette distance que l'on juge la proximité positionnelle.

4.3. Conventions

Convention n° 1 : Mots séparés par ";"

Cette convention est relative à la manière de spécifier les mots qui doivent constituer le contexte positionnel recherché. Les mots en question sont écrits séparés par ";" comme le montre l'écriture formelle suivante :

M1; M2; M3 etc.

L'expression recherchée est constituée par les 3 mots (ou parties de mots) M1, M2 et M3 dont la distance qui les sépare ne doit pas dépasser une valeur fixée (20 caractères par exemple).

Convention n° 2 : Mots interchangeables séparés par ","

Cette convention est relative à la manière de spécifier les mots interchangeables dans les contextes positionnels recherchés. Les mots en question sont écrits séparés par "," comme le montre l'écriture formelle suivante :

M11, M12 ; M21, M22 ; M31, M32, etc.

Les expressions recherchées sont du type précédent, sauf qu'on permet de mettre à la place de M1 des mots qui le représentent (M11, M12) (idem pour M2 et M3).

Convention n° 3 : emploi du symbole "="

Cette convention est relative à la manière de spécifier les mots entiers (explicites) devant faire partie du contexte positionnel recherché. Les mots en question sont écrits précédés par le symbole "=" comme le montre l'écriture formelle suivante :

M1 ; =M2 ; M3 etc.

Les expressions recherchées sont du type de la convention 1 sauf que =M2 spécifie que le mot M2 doit figurer explicitement en entier dans les expressions.

Convention n° 4 : Emploi des racines arabes

Cette convention est relative à l'emploi des racines arabes comme critère de recherche par expression. Les racines impliquées représentent la classe des mots interchangeables (au sens de la convention 2). Cette convention suppose le choix de l'option "Racine arabe". Les racines en question sont écrites séparées par le symbole ";" comme le montre l'écriture formelle suivante :

Rac 1; Rac 2 ; Rac 3 etc.

Les expressions recherchées sont du type de la convention 1 sauf que Rac i spécifie qu'elle représente la suite des mots : Mi1, Mi2, Mi3,… (Mots séparés par virgules et donc interchangeables au sens de la convention 3), qui ont comme racine Rac i.

Convention n° 5 : Emploi des racines de façon individuelle (emploi du symbole "*")

Cette convention est relative à la manière de spécifier l'emploi des racines arabes de façon individuelle en vue de pouvoir utiliser toutes les conventions, selon les besoins, dans une même requête. Les racines en question sont écrites précédées par le symbole "*" comme le montre l'écriture formelle suivante :

***RAC1; Mot2; Mot3** etc.

Les expressions recherchées sont du type de la convention 1 sauf que *RAC1 spécifie qu'elle représente la suite des mots : Mot11, Mot12, Mot13, … (Mots séparés par virgules et donc interchangeables au sens de la convention 3), qui ont comme racine RAC1. Il est important de faire remarquer que l'emploi du symbole "*" suppose obligatoirement la non utilisation de la convention 4 qui régit l'emploi des racines au niveau de tous les éléments de la requête.

Convention n° 6 : Elle est relative à l'affinité entre caractères

Cette convention est relative à l'emploi de l'affinité entre caractères comme critère de recherche par expression. Les éléments impliqués représentent la classe des mots interchangeables (au sens de la convention 2) générés par le mécanisme "classe de caractères". Cette convention suppose le choix de l'option "Classe de caractères". Les éléments en question sont écrits séparés par le symbole ";" comme le montre l'écriture formelle suivante:

Mot 1; Mot 2; Mot 3 etc.

Dans l'exemple suivant, la recherche se fait avec le mot "néttoyer" au lieu de "nettoyer"

Machine ; néttoyer

en recourant au mécanisme "classe de caractères" qui utilise pour cet exemple la classe de caractères suivante :

{e, é, è, ê}

Le schéma des expressions recherchées est le suivant :

*	mots en affinité caractères avec **machine**	*	mots en affinité caractères avec **néttoyer**	*

Cet exemple permet de trouver les expressions du type suivant :

Machine pour nettoyer

Convention n° 7 : Emploi de l'affinité morphologique entre mots

Cette convention est relative à l'emploi de l'affinité morphologique entre mots comme critère de recherche par expression. Les éléments impliqués représentent la classe des mots interchangeables (au sens de la convention 2) générés par le mécanisme "affinité morphologique". Les éléments en question sont écrits séparés par le symbole ";" comme le montre l'écriture formelle suivante :

Mot 1; Mot 2; Mot 3 etc.

Les expressions recherchées sont du type de la convention 1 sauf que Mot i sera remplacé par la base du mot (mot débarrassé du préfixe et du suffixe) sur la base du mécanisme "affinité morphologique".

Remarque : Cette convention concerne le français. Le cas de l'arabe est régi par les conventions 4 et 5 relatives à l'affinité morphologique basée sur l'emploi de la racine des mots arabes.

Dans l'exemple suivant nous allons rechercher les expressions contenant les mots "**Machines**", "**lavage**", en recourant au mécanisme "affinité morphologique" basée sur l'emploi des préfixes et suffixes :

Machines ; lavage

Sur cet exemple, le système calcule les représentants des deux familles de mots :

- Mach → Machine, machines, machin, etc.
- Lav → Laver, lavage, laverie, etc.

Le schéma des expressions recherchées est le suivant :

*	Affinité morphologique avec **machines**	*	Affinité morphologique avec **lavage**	*

Ce qui permet de trouver les expressions du type suivant :

Machine pour le lavage
Machine à laver
Machin de lavage
etc.

Convention n° 8 : Emploi de l'affinité sémantique entre mots

Cette convention est relative à l'emploi de l'affinité sémantique entre mots comme critère de recherche par expression. Les mots impliqués représentent la classe des mots interchangeables. Les mots en question sont écrits séparées par le symbole ";" comme le montre l'écriture formelle suivante :

Mot 1 ; Mot 2 ; Mot 3 etc.

Les expressions recherchées sont du type de la convention 1 sauf que Mot i spécifie qu'elle représente la suite des mots : Mi1, Mi2, Mi3,… (Mots séparés par virgules et donc interchangeables au sens de la convention 3). Ces mots constituent le champ sémantique du mot Mot i. et seront générés par le mécanisme "Affinité sémantique" (consultation du lexique relatif au champs sémantiques métiers).

Dans l'exemple suivant nous allons rechercher les expressions contenant les mots "**machine**", "**laver**", en recourant au mécanisme "affinité sémantique" basée sur l'emploi du lexique des champs sémantiques métiers :

Machine ; laver

Pour cet exemple, le système consulte le lexique relatif aux champs sémantiques métiers pour obtenir les expressions. Voici à titre d'exemple ce qui a été utilisé come relations sémantique :

- Machine → mach , engin , appareil
  → آلة، آلات، مكنة، مكنات، جهاز،أجهز
- Lav → lav, nett..
  → نظف، نظيف، نظوف
  غسل ، غسيل ، غتسال ، غسول ، غاسل ،غوسل .

Le schéma des expressions recherchées est le suivant :

*	Affinité sémantique avec **machine**	*	Affinité sémantique avec **laver**	*

Ce qui permet de trouver les expressions du type suivant dans les trois langues :

Machine à laver
Machine de nettoyage
Engins pour le lavage
Machines pour nettoyer
Machin de lavage
etc.

5. Conclusion

Nous avons présenté dans cet article les outils qui peuvent être utilisés pour faire des recherches dans les textes naturels d'une façon générale et ceux de la langue arabe en particulier. Nous allons rappeler dans ce qui suit l'essentiel des options de recherche qui ont été présentées :

- Recherche par emploi de l'affinité caractère qui permet de faire abstraction de l'accentuation en français ou la confusion entre certaines lettres en arabe;
- Recherche par emploi de l'affinité morphologique qui permet de faire état de la notion de famille de mots;
- Recherche par les racines arabes qui se caractérise par la simplicité d'utilisation et l'obtention de résultats exhaustifs; même si cela demande de supporter un certain bruit du à l'ambiguïté lié au hors contexte;
- Recherche par emploi de l'affinité sémantique qui permet de faire jouer la recherche par concept.
- Recherche par expression qui permet de considérer les options précédentes dans un cadre contextuel positionnel. Nous avons eu l'occasion également de mettre en évidence l'idée du schème syntaxique à l'instar du schème morphologique que connait la langue arabe. Nous fondons beaucoup d'espoir sur cette approche qui peut constituer une méthodologie permettant l'extraction de relations sémantiques dans les textes et les utiliser pour construire un système sémantique permettant de poser des questions relatives aux concepts évoqués dans les corpus utilisés.

Ouverture des Noms de Domaine Internet Génériques par l'ICANN (New Generic Top Level Domains)

Ali Bouallou
ISOC, Maroc
ali.bouallou@gmail.com
Conseil Royal Consultatif des Affaires Sahariennes
ab@corcas.com

Résumé

L'ICANN est une organisation non gouvernementale dont le siège est aux Etats-Unis (Californie) et qui préside à la destinée de l'Internet. Elle prend en charge l'allocation des espaces d'adressage et de nommage sur Internet aussi bien pour les noms génériques (.com, .net, …) que les codes de pays (.ma, .fr, .es, …). Elle a pour mission également de gérer les serveurs racines (root servers) à travers la gestion du système de noms de domaines Internet (Domain Name System).

ICANN a pour rôle aussi de préserver la sécurité, la stabilité et l'interopérabilité de l'Internet dans le respect de la concurrence et de la représentativité des communautés Internet mondiales et ce, dans un esprit de consensus pour le bien de tous. Le slogan de la communauté de l'ICANN est le suivant : we believe in rough consensus and working codes.

1. Introduction

L'ICANN est une organisation non gouvernementale dont le siège est aux Etats-Unis (Californie) et qui préside à la destinée de l'Internet. Elle prend en charge l'allocation des espaces d'adressage et de nommage sur Internet aussi bien pour les noms génériques (.com, .net, …) que les codes de pays (.ma, .fr, .es, …). Elle a pour mission également de gérer les serveurs racines (root servers) à travers la gestion du système de noms de domaines Internet (Domain Name System).

ICANN a pour rôle aussi de préserver la sécurité, la stabilité et l'interopérabilité de l'Internet dans le respect de la concurrence et de la représentativité des communautés Internet mondiales et ce, dans un esprit de consensus pour le bien de tous. Le slogan de la communauté de l'ICANN est le suivant : we believe in rough consensus and working codes.

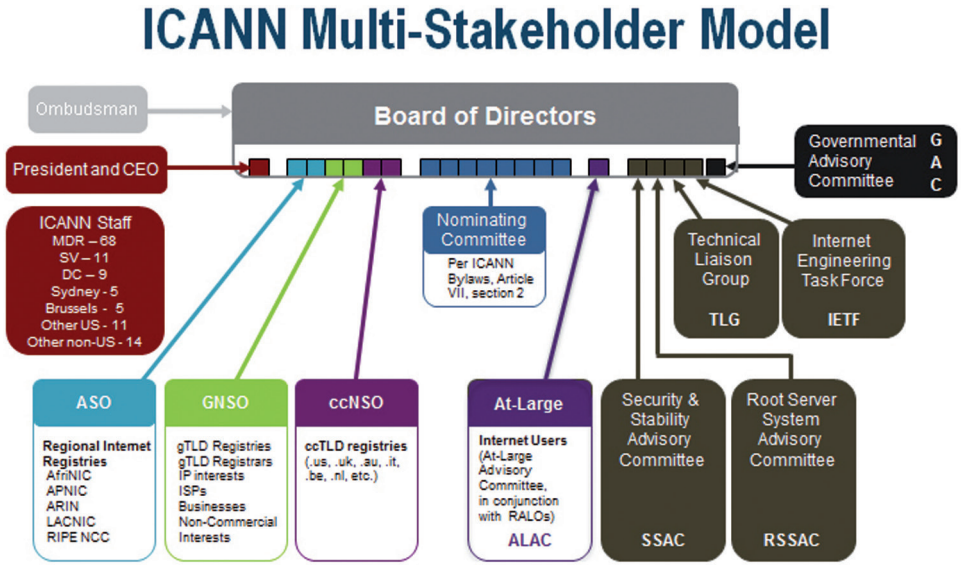
ICANN est composée de plusieurs groupes de travail réunissant professionnels, utilisateurs et société civile, entités gouvernementales et observateurs internationaux désignés. Ces groupes de travail représentent des forces de proposition pour le directoire de l'ICANN. En effet, à la fin de chaque rencontre publique de l'ICANN telle que celle tenue dernièrement à Prague (http://prague44.icann.org), les travaux des groupes de travail sont soumis aux votes des membres du directoire de l'ICANN.

L'ICANN est une organisation « transparente » qui met à la disposition de la communauté Internet, pour commentaire, tous les documents traitant des sujets prévus lors des trois réunions publiques annuelles tenues aux alentours des moins de mars, juin et octobre de chaque année.

Chaque structure composant l'ICANN dispose d'un site web sur lequel sont déposés pour consultation/téléchargement tous les documents de travail.

2. Présentation de l'ICANN

Je vous présente ci-contre les différentes structures composant l'ICANN :



• **Board of Directors -***http://www.icann.org/en/groups/board*

C'est le conseil d'administration de l'ICANN. Il est composé de 21 membres désignés de la manière suivante :

- Le Président Directeur Général nommé par le comité exécutif du Board.
- 8 membres sont proposés par le comité des candidatures (Nominating Committee).
- 2 membres sont proposés par le groupe GNSO.
- 2 membres sont proposés par le groupe ccNSO.
- 1 membre proposé par le groupe At-Large.
- 2 membres proposés par le groupe ASO.
- 1 membre de liaison avec le groupe GAC.
- 1 membre de liaison avec le groupe IETF.

- 1 membre de liaison avec le groupe SSAC.
- 1 membre de liaison avec le groupe RSSAC.
- 1 membre de liaison avec le groupe TLG.

Ces membres siègent au conseil d'administration de l'ICANN pour des mandats renouvelables de 3 ans. Les membres proposés par le comité de candidatures (Nominating Committee) sont ouverts à la communauté Internet mondiale.

• Governmental Advisory Committee (GAC) – *http://gac.icann.org*

Ce comité consultatif réunit tous les représentants des gouvernements. Certains gouvernements sont présents en force et orientent pour leurs intérêts et ceux de leur communauté Internet les documents de travail soumis au Board de l'ICANN. Ce comité a pour rôle de veiller au respect des règles établies par l'ICANN en matière de développement de l'Internet.

• Country Code Names Supporting Organisation (ccNSO) – *http://ccnso.icann.org*

Cette organisation prend en charge la gestion des noms de domaines des pays et territoires selon les règles établies par l'ICANN. Elle se base pour ce faire sur la norme ISO 31661- relative à la liste des pays et territoires.

• At-Large Advisory Committee (ALAC) – *http://alac.icann.org*

At-Large est le nom donné à la communauté des utilisateurs finaux participant au développement de l'Internet.

• Generic Names Supporting Organization (GNSO) – *http://gnso.icann.org*

C'est la plus grande structure d'ICANN en termes d'élaboration des politiques de développement de l'Internet. GNSO s'efforce de maintenir les noms de domaine génériques de premier niveau de manière équitable et ordonnée favorisant l'innovation et la concurrence. Le Maroc n'est pas représenté dans cette structure.

• Address Supporting Organization (ASO) – *http://aso.icann.org*

L'objectif de cette structure est de (re)définir les recommandations en termes d'adressage IP. Elle comprend les organes régionaux d'enregistrement IP à savoir :

- African Network Information Centre (AfriNIC) pour le continent africain.
- Asia Pacific Network Information Centre (APNIC) pour la région Asie-pacifique.
- American Registry for Internet Numbers (ARIN) pour la région d'Amérique du Nord.
- Latin American & Caribbean Network Information Centre (LACNIC) pour la région d'Amérique Latine et des îles Caraïbes.
- Réseaux IP Européens Network Coordination Centre (RIPE NCC) pour la région d'Europe, du Moyen-Orient et une partie de l'Asie Centrale.

• Nominating Committee – *http://nomcom.icann.org*

Ce groupe est en charge de la sélection de 8 membres du conseil d'administration d'ICANN. Il prend en charge également la sélection d'autres ressources conformément aux règlements (bylaws) d'ICANN.

• Security and Stability Advisory Committee – *http://www.icann.org/en/groups/ssac*

L'objectif de ce groupe est d'assister le Board d'ICANN dans la sécurité et l'intégrité des systèmes d'allocation des adresses et des noms de domaine. SSAC est engagé dans l'évaluation des différents risques liés à l'attribution des adresses et des noms de domaine.

• DNS Root Server System Advisory Committee – *http://www.icann.org/en/groups/rssac*

Ce groupe est composé des représentants des organisations responsables de la bonne marche des 13 serveurs de noms racines (DNS root servers).

• Internet Engineering Task Force – *http://www.ietf.org*

Le règlement d'ICANN impose la présence d'un représentant non-votant de l'IETF au sein du Board.

• Technical Liaison Group – *http://www.icann.org/en/groups/tlg*

Ce groupe réunit 4 organisations :

- European Telecommunications Standards Institute (ETSI),
- International Telecommunications Union's Telecommunication Standardization Sector (ITU-T),
- World Wide Web Consortium (W3C), et
- Internet Architecture Board (IAB).

L'objectif de ce groupe est de mettre en relation les membres d'ICANN avec les bonnes ressources dans chaque organisation en fonction des besoins. Il faut noter qu'ICANN lance des projets d'évaluation et de révision de son organisation de manière périodique afin de déterminer si les structures composant ICANN ont lieu de continuer d'exister, si elles sont performantes et les orientations stratégiques pour leur évolution. Elle fait appel pour cela à des consultants indépendants. Chaque groupe précité peut contenir un ou plusieurs groupes de travail en fonction des besoins. L'accès à ces groupes de travail est ouvert à toute la communauté Internet. Tous les participants aux différentes structures composant ICANN sont bénévoles. Il peut arriver que l'on fasse appel à des consultants et experts indépendants pour accompagner la mise à niveau ou l'implémentation de nouvelles directives pour le développement et le maintien de l'Internet.

3. Participation du Maroc à l'ICANN

Le Ministère du Commerce, de l'Industrie et des Nouvelles Technologies représente le Maroc au GAC.

L'ANRT, en tant que gestionnaire délégué du nom de domaine .ma, représente le Maroc dans le groupe ccNSO.

Entre autres associations et entreprises, Moroccan Internet Society (MISOC) représente le Maroc au sein du groupe ALAC. Le fournisseur de services Internet, Genious Communications, le seul ISP accrédité ICANN en Afrique du Nord, participe également aux réunions d'ALAC.

Le Maroc n'est représenté dans aucune autre structure. Il faut rappeler au passage que la participation aux groupes de travail d'ICANN est bénévole et ouverte à toute personne volontaire disposant de compétences pouvant servir le développement de l'Internet.

En ce qui concerne le renforcement de la représentation marocaine dans les différentes instances de l'ICANN, il serait judicieux pour le Maroc que sa participation soit active à tous les niveaux et que le secteur privé (les opérateurs télécoms & Internet) soient engagés aussi bien dans les organes administratifs que les groupes de travail techniques. Il faut marquer la présence marocaine à chaque réunion publique en commentant les documents de travail publiés par les organes de l'ICANN avant la tenue des réunions publiques. Il est impératif que le Maroc dispose dans chaque organe de représentants anglophones afin de défendre les intérêts de la communauté Internet marocaine.

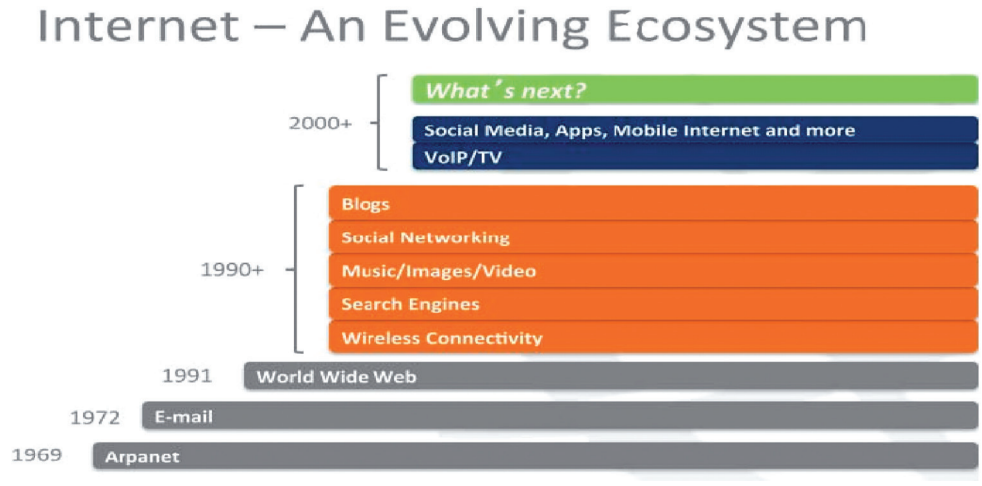
Les dernières réunions tenues d'ICANN ont porté, entre autres sujets, sur les points principaux suivants :

- Les conditions d'introduction des nouveaux noms de domaines génériques (*generic Top Level Domains -gTLDs*) en latin comme les noms de domaines de certaines villes, *.berlin, .nyc* (New-York), *.paris*, *.bcn* (Barcelone)…, de régions comme *.africa* … , de communautés comme *.green, .music*…et de marques déposées comme *.ibm*, *.axa*, …
- Les conditions d'introduction des nouveaux noms de domaines génériques (gTLDs) en arabe ou en tifinaghe (non entamé par le Maroc) ou dans d'autres langues non latines.
- Les conditions d'introduction des noms de domaine de second niveau avec des lettres accentuées, et des noms de domaine composés d'une voire deux lettres.
- Les conditions d'introduction des nouveaux noms de domaines de pays ou de territoires en caractères non latins comme المغرب (*Internationalized Domain Names - IDNs*).
    - Le Maroc, via le Ministère du Commerce de l'Industrie et des Nouvelles Technologies, a répondu positivement à la lettre de l'ICANN envoyée à la communauté Internet en 2009 relative à l'introduction des noms de domaines en arabe en « phase rapide » (*Fast-Track*). Il s'agit d'une procédure préliminaire permettant aux pays prêts à utiliser leur nom de domaine Internet dans leur langue officielle de procéder à sa mise en place. On pourrait faire la même chose pour le script tifinaghe.

- La sécurité des systèmes DNS et les enjeux liés à la sécurisation et à la stabilisation des systèmes DNS installés dans le monde.
- L'introduction de l'adressage IP version 6 (IP v6) permettant de pallier au manque d'adresses IP sur Internet posé par l'adressage IP actuel (IP v4), et de fédérer plusieurs équipements domestiques ou d'utilité publique.
- La relation contractuelle de l'ICANN avec le Département du Commerce Américain et la demande insistante de la communauté Internet mondiale pour que cesse cette dépendance.
- La transparence de la gestion financière de l'ICANN par la mise à disposition de la communauté Internet de tableaux de bord des recettes et dépenses de l'ICANN.

4. Procédure d'attribution des nouveaux noms de domaine génériques (new gTLDs)

4.1. Rappels sur les noms de domaine

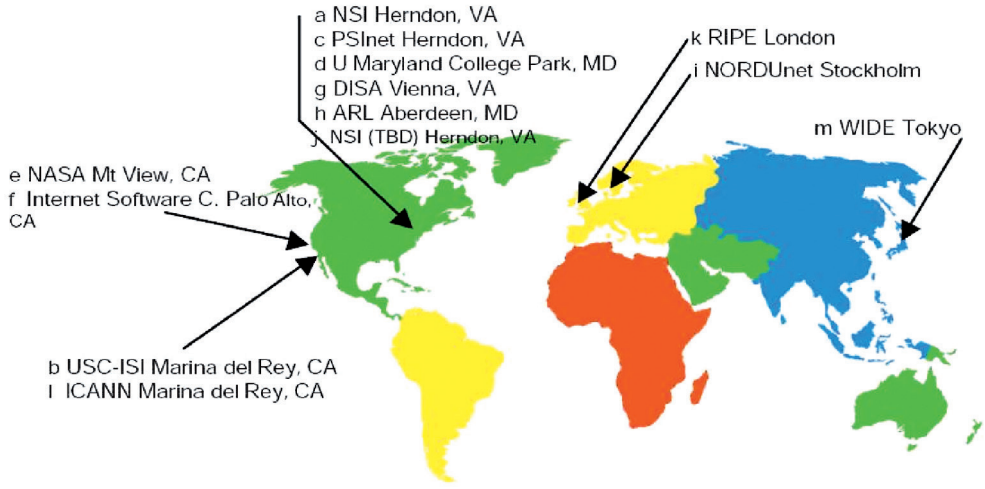


L'Internet est un écosystème ayant évolué de manière ordonnée et stable pour s'imposer dans la vie moderne d'aujourd'hui comme l'outil incontournable d'intégration à la nouvelle économie et le moyen le plus efficient de véhicule du savoir et de la culture. La figure suivante résume l'évolution des services Internet :

L'Internet n'aurait pas pu connaître cette évolution sans le développement du système DNS (Domain Name System) qui consiste en l'association d'une adresse IP unique à chaque site web complètement qualifié sur Internet en d'autres termes disposant d'un nom de domaine composé d'un premier niveau et d'un second voir plusieurs niveaux. Ce service d'association d'adresse IP à un nom complètement qualifié est fourni par des serveurs de noms.

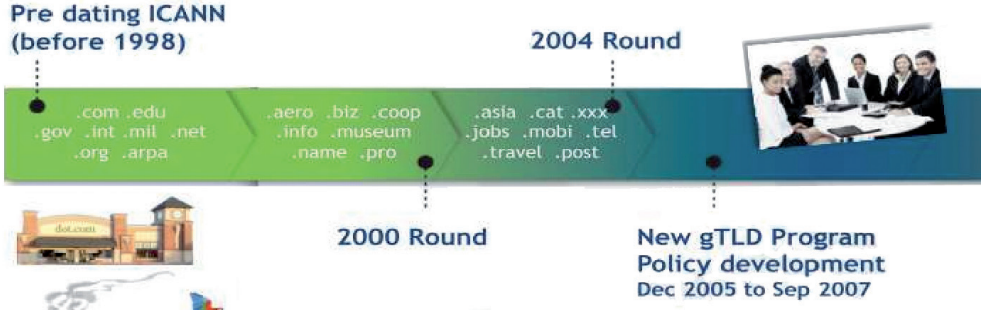
Les serveurs de noms racines, qui sont de 13 et qui sont installés un peu partout dans le monde, sont les serveurs pouvant résoudre l'ensemble des sites web déclarés sur la toile.

La figure suivante présente les emplacements des 13 serveurs root :



L'anatomie d'un nom de domaine est la suivante :

Le premier niveau correspond à l'un des 21 noms de domaine Internet opérationnels maintenus par ICANN. Il s'agit des noms de domaine suivants, .aero, .asia, .biz, .cat, .com, .coop, .info, .int, .jobs, .mobi, .museum, .name, .net, .org, .pro, .tel, .travel, .xxx, .edu, .gov, et enfin .mil. La mise en service de ces noms de domaine a été effectuée selon le calendrier suivant :

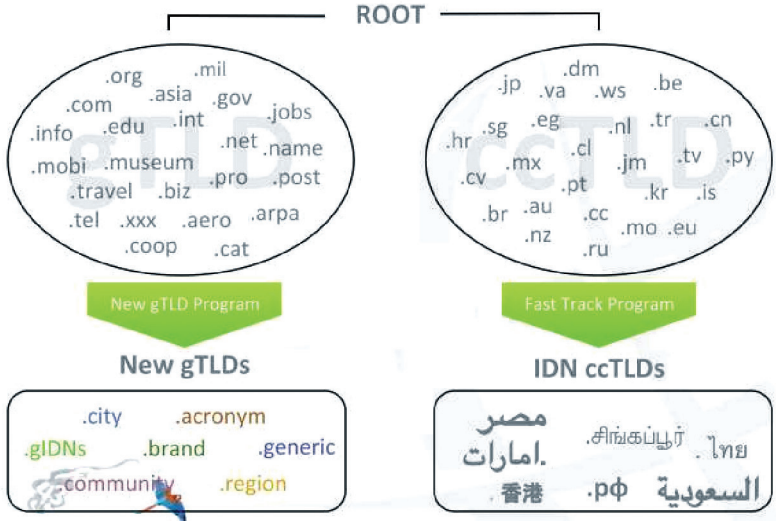


Dans cette figure, on voit bien que les premiers noms de domaine ont été mis en service dès 1998. Un second groupe de noms de domaine a été mis en service en 2000, un autre en 2004.

Entre 2005 et 2007, ICANN a travaillé sur les conditions d'attribution, de manière illimitée, des nouveaux noms de domaine ou extensions (ASCII et IDNs).

Il faut rappeler que cette liste sera renforcée en fonction du résultat des demandes d'attribution des noms de domaine de premier niveau échues le 12 Avril 2012. Le second niveau correspond à toute chaîne de caractères représentant une marque, une communauté, une institution…etc. Le troisième niveau correspond à une segmentation du second niveau en services, départements/régions, localités, etc.

La figure suivante représente la répartition des noms de domaine à partir de la racine d'Internet :



4.2. Approbation de l'ouverture d'Internet sur les nouveaux gTLDs

En juin 2008, le conseil d'administration d'ICANN a approuvé l'ouverture de l'Internet sur de nouveaux noms de domaine génériques. Ce fût un tournant historique dans le développement de l'Internet avec tout ce que cela implique comme défis pour sa sécurité et sa stabilité et ce, malgré les voix qui se sont élevées contre les risques liés, entre autres méfaits, au terrorisme et au respect de la protection intellectuelle.

ICANN, soutenue par la communauté Internet mondiale, a eu les arguments suivants pour défendre cette ouverture historique :

- augmenter le choix et la compétition dans l'espace de noms de domaine.
- créer une plateforme pour l'innovation.
- introduire dans les noms de domaine des chaînes de caractères telles que le chinois, l'arabe, le cyrillique et tout autre langue non-latine.
- augmenter les communautés culturelles, linguistiques et géographiques.
- augmenter le nombre de sites dans les langues minoritaires.

- mettre en avant les noms géographiques.
- permettre aux communautés d'accéder au savoir dans leur langue maternelle.

Après juin 2008, une ébauche du guide de candidature pour l'octroi d'un nom de domaine générique a été établie par la communauté Internet après l'approbation des différentes structures composant ICANN et celle de son conseil d'administration. Le premier draft du guide de candidature a vu le jour en octobre 2008 et il a été approuvé par le conseil d'administration d'ICANN en juin 2011. Le démarrage des soumissions des demandes d'attribution a eu lieu le 12 janvier 2012 et s'est poursuivi jusqu'au 12 avril 2012. La version finale du guide de candidature peut être téléchargée à partir du site d'ICANN, http://newgtlds.icann.org/en/applicants/agb. Tout au long de cette période, des informations mises à jour ont été publiées sur le site d'ICANN concernant le processus de soumission en ligne, les problèmes techniques lors de la soumission des demandes et l'évolution du nombre de demandes.

4.3. Conditions d'attribution d'un gTLD

Toute entité de n'importe où dans le monde qui répond aux critères et pré-requis définis dans le guide de candidature est éligible d'une demande d'attribution d'un gTLD.

La lecture du guide de candidature est primordiale avant toute soumission de demande sur TAS. Une attention particulière devra être portée aux conditions de candidatures des noms de domaine géographiques ainsi que pour les noms de domaine internationalisés (IDNs).

Pour soumettre sa demande, ICANN a développé un système en ligne TAS (TLD Application System) accessible via le site officiel d'ICANN. Le processus de soumission est en anglais. ICANN a mis à la disposition de la communauté un guide d'utilisation de TAS ainsi qu'un ensemble de questions/réponses afin d'en faciliter l'exploitation. Les soumissionnaires sont tenus de respecter les délais imposés par ICANN.

Les frais d'évaluation d'une demande d'attribution de gTLD est de 185.000,00 US$. Ce montant est à verser à ICANN. A cela s'ajoute 5.000,00 US$ de frais d'enregistrement sur TAS. Le remboursement est appliqué dans certains cas. D'autres frais peuvent s'appliquer sans qu'ils soient à reverser à ICANN mais à d'autres intervenants pendant le processus d'évaluation. Une fois accordé, des frais annuels de 25.000,00 US$ relatifs à la gestion du nom de domaine devront être versés à ICANN ainsi que 0.25 US$ pour chaque transaction effectuée (nom de domaine de second niveau vendu ou reconduit). Ces frais n'incluent pas les dépenses relatives au maintien de la plateforme technique et administrative de gestion du nom de domaine que doit assumer pleinement le soumissionnaire.

Dans un souci de servir un public large et de permettre au plus grand nombre d'accéder de manière équitable au programme des nouveaux gTLDs, ICANN a mis en place une procédure d'assistance aux candidats (Applicant Support Program) aussi bien financièrement qu'administrativement et techniquement. ICANN a entamé une large campagne de communication dans les différents médias pour la promotion de ce programme d'assistance aux candidatures.

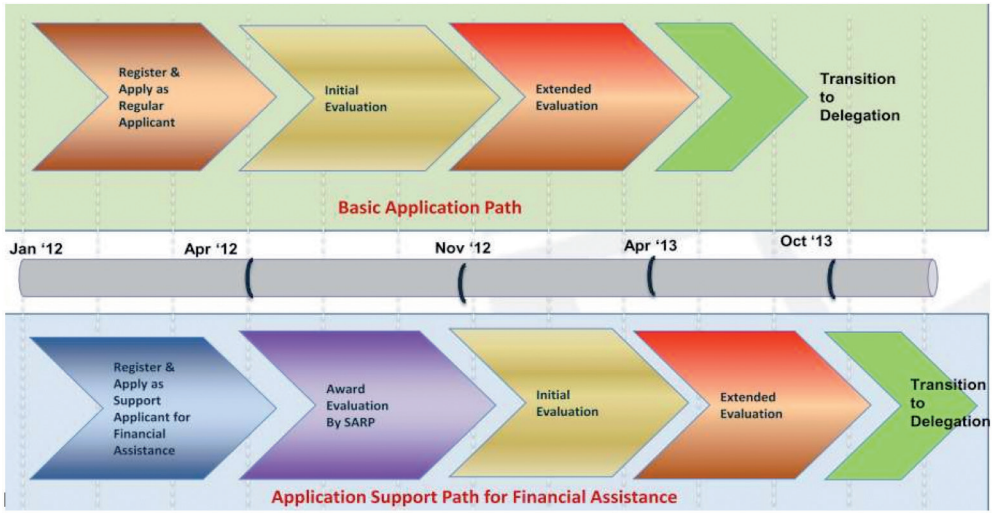
Le conseil d'administration a autorisé la mobilisation de 2.000.000,00 US$ pour cette procédure d'assistance. Elle devrait bénéficier à 14 demandes dans cette première phase.

Pour ce faire, ICANN a défini des critères de sélection relatifs à l'intérêt suscité par le public, aux besoins de financement et à la capacité financière.

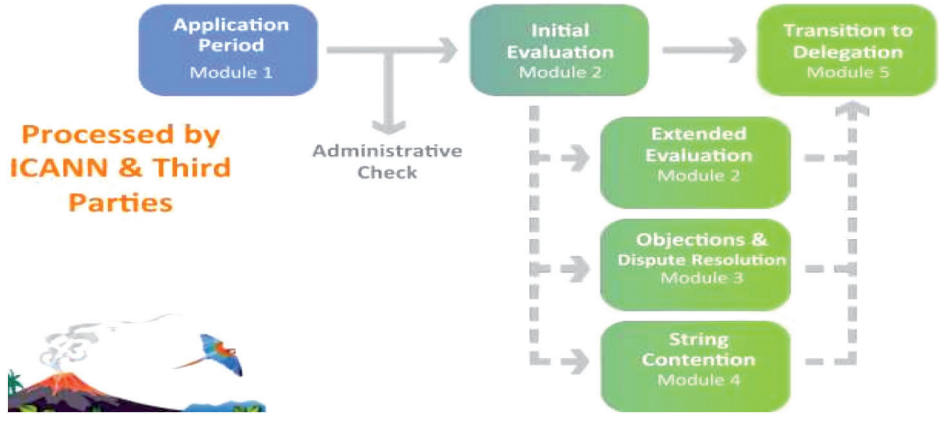
Il faut savoir que lorsqu'une demande d'attribution ne remplit pas les critères susmentionnés, elle est disqualifiée non seulement de la procédure d'assistance financière mais aussi de tout le programme des nouveaux gTLDs.

C'est ainsi que les candidats retenus pour cette assistance peuvent bénéficier d'un rabais de 47.000,00 US$ ou de paiements échelonnés.

La figure suivante présente les étapes d'évaluation des demandes d'attribution dans le cadre du programme d'assistance aux candidatures :



La figure suivante présente de manière simplifiée les étapes de traitement d'une soumission de demande d'attribution d'un gTLD :



Après la période arrêtée pour la soumission des candidatures, ICANN procède au lancement de la procédure d'évaluation des demandes par la vérification de l'exhaustivité des données administratives. Une évaluation initiale est alors entamée en prenant en compte les dépôts d'objection selon les mécanismes arrêtés dans le guide de candidature. Lorsque cela s'avère nécessaire, une évaluation approfondie est effectuée avant la transition vers une délégation du nom de domaine demandé.

Celui-ci peut faire l'objet de litiges par l'entremise d'autres demandes d'attribution du même nom de domaine ou par objection d'une communauté habilitée ou à cause d'un conflit de chaînes ou tout simplement lorsque les évaluateurs estiment que le nom de domaine aura un intérêt limité auprès des internautes.

Les figures suivantes présentent le déroulement des différentes phases d'évaluation des demandes :

2012	29 March 12 April	TAS Registration Closes Application Window Closes
	1 May	Strings Posted Opens: ✓ Objection period ✓ Application Comments period ✓ GAC Early Warning ✓ GAC Advice period
	12 June	Initial Evaluation beings
	30 June	Close of Application Comments period Close of GAC Early Warning period
2012	12 November	Close of Initial Evaluation → Results posted
	29 November	Last day to elect Extended Evaluation
	1 December	Last day to file an Objection Close of GAC Advice period Begins: ✓ Extended Evaluation ✓ Transition to Delegation (for Clean Applications) ✓ String Contention (for Applications not in Extended Eval or Dispute Resolution)
2013	30 April	Close of Extended Evaluation Close of Dispute Resolution Results & Summaries Posted
	15 May	String Contention opens (for Applications w/ variables)
	30 May	String Contention closes (for Clean Applications → Results posted

Ce chronogramme peut faire l'objet de mise à jour à cause des problèmes techniques survenus au lancement de TAS.

Il faut noter au passage que l'avis du GAC doit être présenté avant la fin de la période de dépôt d'objections.

4.4. Récapitulatif des demandes d'attribution de nouveaux gTLDs

Depuis le 12 janvier 2012, ICANN a reçu 1930 demandes de 60 pays des 5 régions définies par ICANN :

- 17 demandes de la région Afrique (0,9% des demandes).
- 303 demandes de la région Asie-Pacifique (15,7%).
- 675 demandes de la région Europe (35%).
- 24 demandes de la région Amérique Latine (24%).
- 911 demandes de la région Amérique du Nord (47,2%).

Il faut savoir aussi que :

- 230 demandes ont été effectuées par plus d'un soumissionnaire.
- 751 demandes sont sujettes à des conflits de chaînes.
- 116 demandes concernent des noms de domaines internationalisés (IDNs).

Les extensions qui présentent les conflits de chaînes sont les suivantes, .app (13), . home (11), .inc (11), art (10), blog (9), book (9), llc (9), .shop (9), .design (8), .cloud (7), .hotel (7), .love (7), .ltd (7), .mail (7), .news (7), .store (7), .web (7).

Trois soumissionnaires ont fait appel à la procédure d'assistance ou d'aide aux candidatures, .kids, .idn et .ummah.

Il faut noter au passage que la période de commentaires et d'objections est définie comme suit et reste ouverte à la communauté Internet mondiale via le site d'ICANN :

- Les commentaires sont recevables pendant 60 jours du 13 juin 2012 au 12 août 2012.
- Les objections officielles sont recevables pendant 7 mois jusqu'au 12 février 2013.

La liste complète des demandes d'attribution des nouveaux gTLDs est disponible sur le lien suivant :
http://newgtlds.icann.org/en/program-status/application-results/strings-1200utc13jun12-en.

5. Enjeux pour le Maroc par rapport à l'ouverture de l'Internet sur les nouveaux gTLDs

A l'instar des autres pays intégrés au système économique mondial, le Maroc est concerné par l'ouverture de l'Internet sur les nouveaux gTLDs. En effet, les grandes entreprises marocaines ainsi que tout patrimoine national culturel, touristique et géographique peuvent prétendre à un nouveau gTLD de type .marrakech par exemple.

Comme les demandes d'attribution d'un nouveau gTLD ne sont pas nécessairement liées au lieu de résidence, un membre de la communauté Internet mondiale peut prétendre à un gTLD représentant un joyau national. Pour éviter de réagir en aval dans le cadre de procédures d'objection, le Maroc devrait anticiper les éventuelles demandes portant atteinte aux acquis nationaux et ce, dans les scripts latins et non latins (y compris arabe & tifinaghe).

Pour ce faire, le Maroc se doit de renforcer sa présence au sein d'ICANN afin de défendre au mieux les intérêts des entreprises et institutions marocaines à l'échelle mondiale. Une instance nationale devra se charger d'une veille informationnelle pour le suivi des demandes d'attribution de nouveaux gTLDs et d'élaborer les éventuelles objections qui s'imposent.

Système de Recherche d'un Mot-Image Cas des Mots Arabes Imprimés

Anass Smaili Ali Lasfar Mohamed Sbihi
LASTIMI, EST Salé, Maroc
{anass.smaili, mohamed.sbihi}@yahoo.Fr
ali.lasfar@gmail.com

Résumé

La technique de la reconnaissance optique de caractère présente un intérêt de recherche particulier, que les laboratoires ne cessent d'améliorer et de développer. Cette technique constitue par excellence, une alternative importante à la saisie du texte. Le fondement de la langue arabe, nous a incité à déployer de grands efforts afin de réussir à maîtriser la reconnaissance optique de ces caractères, sinon, il s'avère nécessaire de concevoir d'autres démarches, par conséquent une approche basée sur la recherche d'un mot-image dans un texte-image.

L'objectif de ce travail est de réaliser des recherches mot-image dans un texte-image du bulletin officiel marocain.

1. Introduction

Les administrations publiques possèdent des fonds documentaires colossaux. Dès lors, l'exploitation de ces fonds pour alimenter les bases de données nécessite, dans ce cas, une opération de numérisation et de conversion en texte ou bien une saisie sur ordinateur, une tâche qui s'avère très fastidieuse et demande une équipe spécialiste en traitement de texte.

Sur le marché, il y a des outils destinés pour la Reconnaissance Optique du Caractère Arabe (ROCA). Ces logiciels offrent souvent un résultat satisfaisant mais pas dans tous les cas. Cela nécessite par la suite une vérification et une revue totale du résultat de cette opération de reconnaissance.

Nous allons commencer par présenter la problématique, un aperçu sur la langue arabe, les travailles sur la ROCA, les travaux connexes et ensuite le système que nous proposons pour la recherche du mot-image. En conclusion, nous proposerons les perspectives relatives à ce genre de recherche.

2. Problématique

Le bulletin officiel, édition générale, et le document officiel édité par l'Etat qui regroupe tous les textes législatifs et règlementaires en langue arabe. La numérisation de ce dernier en format PDF, depuis 1913, a constitué un facteur encourageant à exploiter ce fond

documentaire juridique. L'utilisation d'un système de reconnaissance optique de caractère, version qui traite la langue arabe, n'a pas donné des résultats très satisfaisants (Figure 1), ce qui a empêché d'aller plus loin dans la mise en valeur de ce patrimoine documentaire juridique.

b a

Figure 1 : (a) texte-image extraite du bulletin officiel (b) résultat de la reconnaissance

Néanmoins, nombreuses sont les demandes d'utilisateurs qui souhaitent juste avoir où se situe l'information juridique dans les différents numéros du bulletin officiel. De notre côté, nous avons concentré notre effort sur ce besoin en proposant un système qui va donner à l'utilisateur le moyen de rechercher dans le texte-image. Le texte-image n'est qu'une page numérisée du bulletin officiel (Figure 2).

Le contexte de cet article est de proposer un système qui, à travers une base de données contenant les différentes pages du bulletin officiel en arabe, va rendre possible la recherche d'un mot-image dans un texte-image.

مرسوم رقم 2.95.610 صادر في 8 جمادى الأولى 1416 (4 أكتوبر 1995) بالموافقة على مقرر مجلس جماعة سلا – بطانة بعمالة سلا القاضي بتفويت هذه الجماعة قطعتين أرضيتين من الأملاك الجماعية الخاصة لفائدة الدولة (الأملاك الخاصة).

الوزير الأول ،

بناء على الظهير الشريف رقم 1.76.583 الصادر في 5 شوال 1396 (30 سبتمبر 1976) المعتبر بمثابة قانون يتعلق بالتنظيم الجماعي ، كما وقع تغييره أو تتميمه ؛

وعلى الظهير الشريف الصادر في 17 من صفر 1340 (19 أكتوبر 1921) في شأن الأملاك البلدية ، كما وقع تغييره أو تتميمه ؛

وعلى القرار الصادر في فاتح جمادى الأولى 1340 (31 ديسمبر 1921) بتحديد كيفية إدارة شؤون الأملاك البلدية ، كما وقع تغييره أو تتميمه ؛

وبعد الاطلاع على مقرر مجلس جماعة سلا – بطانة خلال دورته العادية المنعقدة بتاريخ 28 من ربيع الأول 1415 (6 سبتمبر 1994) ؛

وباقتراح من وزير الدولة في الداخلية وبعد استشارة وزير المالية والاستثمارات الخارجية ،

رسم ما يلي :

المادة الأولى

يوافق على المقرر الصادر عن مجلس جماعة سلا – بطانة في 28 من ربيع الأول 1415 (6 سبتمبر 1994) بتفويت هذه الجماعة قطعتين أرضيتين من الأملاك الجماعية الخاصة موضوعا الرسمين العقاريين رقمي 36020 ر و 31955 ر لفائدة الدولة (الأملاك الخاصة) مساحتهما على التوالي أربعة آلاف وستة وعشرون مترا مربعا (4026 م²) وألف ومائة وتسعة وثلاثون مترا مربعا (1139 م²) تقعان بسلا – بطانة.

وقد رسمت حدود القطعتين الأرضيتين بوضوح في المخطط المضاف إلى أصل هذا المرسوم.

المادة الثانية

ينجز التفويت الموافق عليه بموجب هذا المرسوم بثمن إجمالي قدره مليونان وخمسمائة واثنان وثمانون ألفا وخمسمائة درهم (2.582.500 د) أي على أساس خمسمائة درهم للمتر المربع الواحد (500 د/م²).

المادة الثالثة

يسند إلى رئيس مجلس جماعة سلا – بطانة تنفيذ ما جاء في هذا المرسوم الذي ينشر بالجريدة الرسمية.

وحرر بالرباط في 8 جمادى الأولى 1416 (4 أكتوبر 1995).

الامضاء : عبد اللطيف الفيلالي.

وقعه بالعطف :

وزير الدولة في الداخلية ،

الامضاء : إدريس البصري.

قرار لوزير الفلاحة والاستثمار الفلاحي رقم 1732.95 صادر في 23 من محرم 1416 (22 يونيو 1995) بالموافقة على البرنامج الخاص بإنجاز أشغال التجهيز الداخلي للقطاعين الثاني والثالث من الدائرة السقوية لسبو الأوسط.

وزير الفلاحة والاستثمار الفلاحي ،

بناء على الظهير الشريف رقم 1.69.25 الصادر في 10 جمادى الأولى 1389 (25 يوليو 1969) المعتبر بمثابة قانون الاستثمارات الفلاحية ولاسيما الفصل 13 منه ؛

وعلى المرسوم رقم 2.84.320 الصادر في 22 من شعبان 1404 (24 ماي 1984) القاضي بتحديد القطاعين الثاني والثالث للمنطقة المسقية بحوض سبو المتوسط (أقاليم فاس وتاونات وسيدي قاسم) الخاضعة لأحكام الظهير الشريف المشار إليه أعلاه ؛

وبعد الاطلاع على محضر اللجن المحلية للاستثمار الفلاحي لعمالة زواغة مولاي يعقوب وإقليمي سيدي قاسم وتاونات خلال اجتماعاتها المنعقدة على التوالي بتاريخ 24 يونيو 1994 و 12 و 28 يوليو 1994 ؛

وبعد استشارة وزير الدولة في الداخلية ،

قرر ما يلي :

المادة الأولى

يشمل برنامج سنوات 1994–1995 و 1995–1996 و 1996–1997 الخاص بإنجاز أشغال التجهيز الداخلي القطاعين الثاني والثالث البالغة مساحتهما 6.578 هكتارا الدائرة السقوية لحوض سبو الأوسط بولاية فاس وإقليمي سيدي قاسم وتاونات.

ويتناول البرنامج إنجاز أشغال تتمثل في قلب التربة وتسوية الأرض.

ويمكن اتخاذ التدابير الآتية لانجاز الأشغال المذكورة :

- منع عمليات الحرث والسقي ؛
- إتلاف المزروعات الموجودة.

المادة الثانية

ينشر هذا القرار بالجريدة الرسمية.

وحرر بالرباط في 23 من محرم 1416 (22 يونيو 1995).

الامضاء : حمن أبو أيوب.

قرار لوزير الاسكان رقم 1852.95 صادر في 4 صفر 1416 (3 يوليو 1995) بتفويض الامضاء

وزير الاسكان ،

بناء على الظهير الشريف رقم 1.95.40 الصادر في 27 من رمضان 1415 (27 فبراير 1995) بتعيين أعضاء الحكومة ؛

وعلى الظهير الشريف رقم 1.57.068 الصادر في 9 رمضان 1376 (10 أبريل 1957) في شأن تفويض إمضاء الوزراء وكتاب الدولة ونواب كتاب الدولة كما وقع تغييره وتتميمه ولاسيما الفصلين الأول والثاني منه ؛

Figure 2 : Page extraite du bulletin officiel numéro 4330

3. Langue arabe

La langue arabe est construite à partir d'un alphabet de 28 lettres. La particularité de ses caractères est que chacun d'eux peut avoir quatre formes différentes selon sa position dans le mot (Amin, 1997) (Figure 3).

La diversité de ses polices ainsi que la multitude de ses styles d'écriture sont parmi les critères qui enrichissent la langue arabe mais rendent difficiles le processus de la reconnaissance de caractère.

Pour le bulletin officiel, édition générale, nous avons constaté que la police n'a pas trop changé et a gardé à peu près le même style d'écriture.

	isolated (i)	end (e)	middle (m)	beginning (b)
alif	أ	ـا	ـا	ا
ba	ب	ب	ـبـ	بـ
ta	ت	ت	ـتـ	تـ
tha	ث	ث	ـثـ	ثـ
jim	ج	ج	ـجـ	جـ
ha	ح	ح	ـحـ	حـ
kha	خ	خ	ـخـ	خـ
dal	د	د	ـد	د
dhal	ذ	ذ	ـذ	ذ
ra	ر	ر	ـر	ر
zin	ز	ز	ـز	ز
siin	س	س	ـسـ	سـ
shin	ش	ش	ـشـ	شـ
sadd	ص	ص	ـصـ	صـ
dad	ض	ض	ـضـ	ضـ
tahn	ط	ط	ـطـ	طـ
zah	ظ	ظ	ـظـ	ظـ
ayn	ع	ع	ـعـ	عـ
ghayn	غ	غ	ـغـ	غـ
fa	ف	ف	ـفـ	فـ
qaf	ق	ق	ـقـ	قـ
kaf	ك	ك	ـكـ	كـ
lam	ل	ل	ـلـ	لـ
miim	م	م	ـمـ	مـ
noon	ن	ن	ـنـ	نـ
ha	ه	ه	ـهـ	هـ
waw	و	و	ـو	و
ya	ي	ي	ـيـ	يـ
lamalif	لا	لا	ـلا	لا
tamarbot	ة	ة		

Figure 3 : Les différentes formes des caractères arabes

4. Reconnaissance optique du caractère arabe

La difficulté de la langue arabe (Amin, 1997; Argner, 2008; Chan, 2006; Khan, 2010; Märgner, 2009) nous a conduit à voir les différentes études et les recherches sur la ROCA.

Nous avons commencé par voir les travaux de reconnaissance de caractère qui ne traite pas la langue arabe. (Govindan et Shivaprasad, 1990) ont fait un examen sur les méthodologies de la reconnaissance de caractère.

(Al-Badr et Mahmoud, 1995) ont fait l'étude sur la ROCA, en précisant les difficultés que rencontrent les chercheurs pour la reconnaissance du caractère arabe.

Une étude sur des systèmes de ROCA (Jumari et Ali, 2002; Kanungo *et al*., 1998) est faite afin de déterminer le problème fondamental de la bonne réussite de cette opération qui est la segmentation liée à la structure des phrases de la langue arabe (Zeki et Zakaria, 2008). Un système de reconnaissance de caractère pour la langue arabe est proposé par (Cheung *et al*., 1998).

(Ahmed et Al-Ohali, 2000) viennent en 2000 pour présenter une étude sur l'état d'avancement de la reconnaissance du caractère arabe ainsi que les défis à relever pour pouvoir arriver à un taux de satisfaction plus élevé.

(Khorsheed, 2002) nous donne un examen sur la spécificité de la langue arabe face aux systèmes de reconnaissance. (Lorigo et Govindaraju, 2006) ont examiné de près l'écriture arabe en précisant le sens de la recherche dans ce domaine afin d'améliorer la reconnaissance.

(Shaaban, 2008) a aussi contribué à travers une nouvelle approche qui aide à faire la reconnaissance de caractère en se basant sur les réseaux de neurones.

(Al-Shatnawi *et al*., 2011) ont donné un aperçu général sur le système de reconnaissance de caractère arabe en signalant qu'une phase d'étude de la nature du texte arabe est indispensable pour commencer à concevoir un système de reconnaissance.

De nouvelles approches pour la ROCA basées sur des mesures de dissimilarité calculée sur la base de certains attributs polygonaux sont développées (Chaker et Benslimane, 2011). Pour augmenter le degré de précision de la reconnaissance et aboutir à un résultat plus sûr.

5. Travaux connexes

La section précédente nous a montré les différents travaux de recherche sur la ROCA. (Ataer et Duygulu, 2007) ont réalisé la faiblesse des résultats des systèmes de la reconnaissance optique de caractère et ils ont réfléchi à utiliser le mot-image pour indexer des documents pour les caractères ottomans.

(Yue et Chewlim, 2004) ont fait la même chose pour les documents en langue chinoise afin de faire la recherche d'un mot-image dans un document numérisé.

(Andreev et Kirov, 2008) ont utilisé la distance de Hausdorff afin de pouvoir rechercher le mot-image dans un texte-image pour les caractères bulgares.

6. Correspondance mot-image

Parmi les étapes les plus importantes dans ce système est la partie correspondance, c'est-à-dire faire correspondre le mot-image au texte-image. (Andreev et Kirov, 2008) se sont basés sur la distance de hausdorff (Andreev et Kirov, 2006) en la modifiant afin d'assurer cette tâche. (Rath et Manmatha, 2003) utilisent la déformation temporelle dynamique (DTD), méthode qui calcule la similarité entre deux suites. Les deux chercheurs montrent la force de cet algorithme pour la correspondance de deux mots-image. Dans le même sens, (Meshesha et Jawahar, 2003) à travers une étude sur la correspondance mot-image ont montré aussi la pertinence des résultats de la DTD.

7. Système proposé

Avec une langue aussi diversifiée comme la langue arabe, nous avons opté pour concevoir un système qui va permettre la recherche avec indexation (Eldos, 2003; Ibrahim *et al*., 2005) dans le bulletin officiel (Figure 4) en arabe. Ce système va permettre de parvenir à indexer progressivement les documents PDF numérisés du bulletin officiel.

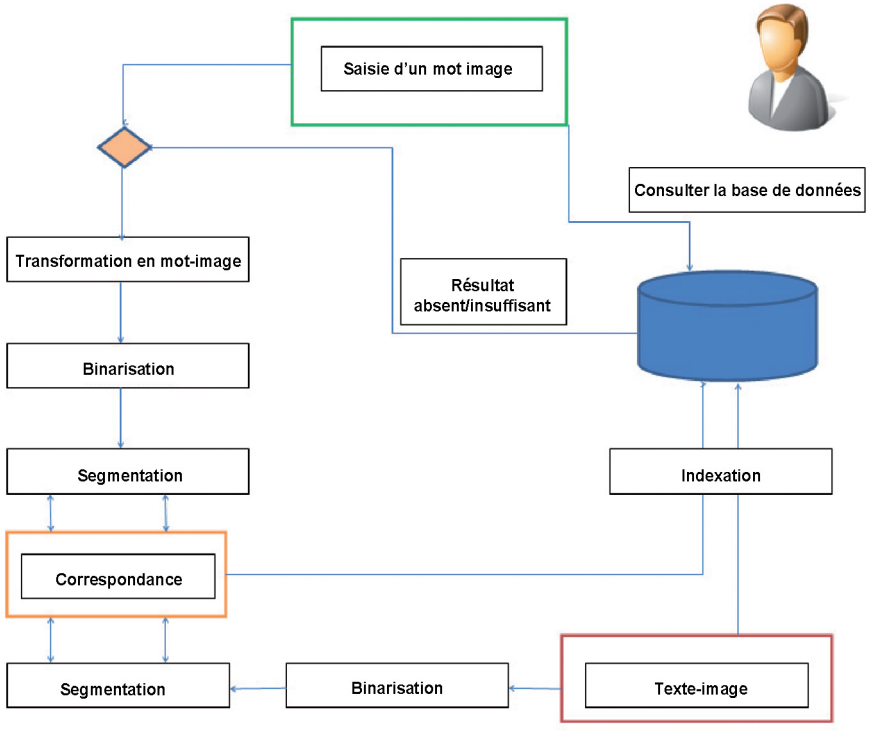


Figure 4: Système de recherche d'un mot-image

Le futur utilisateur de ce système va être en mesure de saisir un mot qui va consulter la base de données en premier lieu, si le résultat n'est pas satisfaisant ou le mot recherché n'est pas encore indexé, le mot sera transformer en mot-image qui va connaître une opération de binarisation (0,1) ainsi qu'un processus de correspondance sera lancer pour trouver le mot-image dans le texte-image.

8. Conclusion et perspectives

Nous avons présenté les différents travaux sur la reconnaissance optique de caractère arabe ainsi que les difficultés que rencontrent les chercheurs pour établir un bon système de reconnaissance. Le problème de la segmentation en caractère est le point focal pour bien réussir cette opération.

Nous avons essayé de proposer un système qui va donner la main à l'utilisateur pour remplir et faire la recherche dans la base de données des texte-image du bulletin officiel version arabe.

Actuellement, nous avons pu segmenter une page du bulletin officiel et à faire correspondre le mot-image au texte-image pour aboutir à un résultat de la recherche. La partie correspondance va être plus détaillé et fera l'objet d'une prochaine publication qui traite les différents algorithmes que nous avons étudié ainsi que notre choix.

Références

Ahmed P., Al-ohali Y. (2000). Arabic character recognition: Progress & challenges. *Journal King Saud Univ.,* vol. 12. pp. 85-116.

Al-Badr B., Mahmoud S. (1995), Survey and bibliography of Arabic optical text recognition, Signal Process. vol. 41. pp. 49-77.

Al-Shatnawi A. M., Al-Zawaideh F. H., Al-Salaimeh S., Omar K. (2011). Offline Arabic Text Recognition – An Overview. *World of Computer Science and Information Technology Journal. 1(5):* 184-192.

Amin A. (1997). Arabic character recognition, In H. Bunke and P. S. P. Wang, editors, Handbook of Character Recognition and Document Image Analysis, chapter 15. World Scientic. pp. 397-420.

Andreev A., Kirov N. (2006). Word image matching in Bulgarian historical documents, Review of the National Center for Digitalization, vol. 8.

Andreev A., Kirov N. (2008). Some variants of Hausdorff distance for word matching. Review of the National Center for Digitization, vol. 12. pp. 3-8.

Argner V., El Abed H. (2008). Databases and Competitions: Strategies to Improve Arabic Recognition Systems. Conference on Arabic and Chinese handwriting recognition pp. 82-103.

Ataer E., Duygulu P. (2007). Matching ottoman words: an image retrieval approach to historical document indexing. *The 6th ACM international conference on image and video retrieval*, ACM, NY, USA, pp. 341- 347.

Chaker I., Benslimane R. (2011). Nouvelle approche pour la reconnaissance des caractères arabes imprimés. Revue méditerranéenne des Télécommunications, 1(2): 87-92.

Chan J., Ziftci C., Forsyth D. (2006). Searching off-line arabic documents. *IEEE Conference on Computer Vision and Pattern Recognition*.

Cheung A., Bennamoun M., Bergmann N.W. (1998). A recognition-based Arabic optical character recognition system. IEEE International Conference on Systems, Man and Cybernetics. pp. 4189-4194.

Govindan V. K., Shivaprasad A. P. (1990). Character recognition - A review. Pattern Recognition, 23(7): 671-683.

Ibrahim K., Eltalmas A., Abubaker M. (2005). *On searching Arabic records in electronic libraries*. International Cataloguing and Bibliographic Control, 34(2): 23-26.

Jawahar C. V., Kumar M. N. S. S. K. P., Kiran S. S. R. (2003). Bilingual OCR for Hindi-Telugu Documents and its Applications. International Conference on Document Analysis and Recognition. pp. 408-412.

Jumari K., Ali M. A. (2002). A Survey And Comparative Evaluation Of Selected Off-Line Arabic Handwritten Character Recognition Systems. *Jurnal Teknologi*, vol. 36. pp. 1-18.

Eldos T. (2003). Arabic Text Data Mining: A Root-Based Hierarchical Indexing Model. *International Journal of Modelling and Simulation*, vol. 23. pp. 158-166.

Kanungo, T., Marton G. E., Bulbul O. (1998). Performance Evaluation of Two Arabic OCR Products. Proc. Of AIPR Workshop on Adv. in Comp. Assist. Recognition, SPIE, vol. 3584. Washington DC, USA.

Khan B., Alghathbar K. S., Khan M. K., AlKelabi A. M., AlAjaji A. (2010). Using Arabic CAPTCHA for Cyber Security. Computer and Information Science, vol. 122. pp. 8-17.

Khorsheed M. S. (2002). *Off-Line Arabic Character Recognition* – A Review. Pattern Analysis & Applications, 5 (1): 31-45.

Lorigo L. M., Govindaraju V. (2006). Offline Arabic handwriting recognition: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5): 712-724.

Märgner V., El Abed H. (2009). Arabic Word and Text recognition – Current Developments. The 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Shaaban Z. (2008). A New Recognition Scheme for Machine-Printed Arabic Texts based on Neural Networks. World Academy of Science, Engineering and Technology, vol. 31. Vienna, Austria.

Rath T., Manmatha R. (2003). Word image matching using dynamic time warping. CVPR, vol. 2, pp. 521-527.

Yue L., Chewlim T. (2004). Chinese word searching in imaged documents. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(2): 229-246.

Zeki A. M., Zakaria M. S. (2008). Challenges in Recognizing Arabic Characters. International Symposium of Information Technology.

Enrichissement Sémantique des Requêtes Multi-Mots

Mohamed Rachdi El Habib Benlahmar El Hassan Labriji
Faculté des Sciences BenM'sik Casablanca, Maroc
{mohamed.rachdi,labriji}@yahoo.fr
h.Benlahmer@gmail.com

1. Résumé

L'expression du besoin de l'utilisateur (requête utilisateur) constitue le point de départ d'un ensemble de processus se terminant par la sélection des documents retournés vers l'utilisateur.

Trouver un document pertinent, c'est tout d'abord bien exprimer le besoin, c'est bien le formuler car c'est une des clefs pour l'amélioration des performances des systèmes de recherche d'information. De ce fait, on constate que la reformulation des requêtes est un processus assez important que ce soit pour le SRI, car il facilite la tâche de la recherche des documents, ou bien pour l'utilisateur final car il permet de retourner un résultat pertinent.

Ce travail propose une technique de reformulation des requêtes composées de plusieurs mots en se basant sur les phrases de définition issues directement d'Internet.

Dans notre proposition, pour chaque mot de la requête initiale, on cherche l'ensemble des phrases de définition qui lui sont correspondantes, et chacune des phrases contient un ensemble de mots appelés dans notre approche « mots associés », l'intersection entre les différents ensembles des mots associés nous donne un seul ensemble des mots communs entre les différentes listes, ces mots qui seront utilisés pour enrichir la requête initiale de l'utilisateur.

Pour tester notre approche, on a fait une comparaison entre le résultat retourné sans reformulation et en utilisant notre approche de reformulation dans Google. Le résultat ainsi obtenu a montré l'apport de notre proposition.

2. Introduction

L'objectif des SRI est la sélection des documents pertinents pour une requête donnée. Ils se basent sur des techniques et algorithmes bien définis, parmi ces techniques on trouve la reformulation des requêtes, c'est une étape essentielle du processus général de la recherche d'information dans les moteurs de recherche modernes. Cette étape consiste a représenter la requête sous une forme facilement compréhensible par l'outil de recherche d'information et de trouver le document pertinent.

La finalité de notre approche est de proposer une nouvelle technique de reformulation (enrichissement) des requêtes en se basant sur les phrases de définition, on utilise ces derniers

pour extraire les mots dits «associés » et parmi ces mots on applique un algorithme pour trouver des mots qu'on appelle « mots communs », ils seront utilisés dans notre cas pour l'enrichissement de la requête.

Dans cet article on commence par un état de l'art sur les différentes techniques de reformulation de requêtes, puis on passe à la présentation de notre approche, une fois achevée on fera une comparaison pour évaluer notre système et on terminera par une conclusion et des perspectives.

3. Etat de l'art

3.1. La réinjection par pertinence

Elle consiste à sélectionner les termes importants appartenant aux documents jugés pertinents par l'utilisateur, et de renforcer l'importance de ces termes dans la nouvelle formulation de la requête (Boughanem *et al*., 1999 ; Ruthven et Lalmas, 2003). Cette méthode a pour double avantage une simplicité d'exécution pour l'utilisateur qui ne s'occupe pas des détails de la reformulation, et un meilleur contrôle du processus de recherche en augmentant le poids des termes importants et en diminuant celui des termes non importants.

3.2. La reformulation par réinjection de pertinence automatique pour la langue arabe

Au début la requête de l'utilisateur est décortiquée en plusieurs mots, pour chaque mot un module de racinisation génère une liste contenant la ou les racines possibles du mot, puis cette liste est envoyée a un autre module appelé module de génération légère qui génère un ensemble de mots à partir de chaque racine. Après, ce module envoie une liste des mots dérivés au module de reformulation, ce dernier reformule la requête en ajoutant tous les mots de la liste dérivée. Ensuite, cette requête intermédiaire est envoyée a un méta-moteur de recherche. Une réinjection de pertinence automatique sera appliquée à N résultats renvoyés par ce méta moteur de recherche, et la nouvelle requête étendue finale sera construite et envoyée au méta moteur de recherche, le résultat retourné sera considéré définitif (El Younoussi *et al*., 2010 ; Benlahmar *et al*., 2010).

3.5. Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées

Cette méthode utilise des ontologies personnalisées pour étendre les requêtes saisies par l'utilisateur (Aimé *et al*., 2010). Elle se base sur des gradients de prototypicalité (représentées, pour la prototypicalité conceptuelle, par des pondérations sur les liens hiérarchiques et les propriétés, pour la prototypicalité lexicale, par des pondérations sur les termes), afin de personnaliser à un utilisateur donné l'extension de la requête saisie ainsi que la quantité des résultats fournis. Cette approche propose donc de faire des ontologies qui constituent elles-mêmes le support de la personnalisation du SI, en ce sens qu'elles représentent un fond cognitif commun à tous les utilisateurs potentiels du système, et qu'il est possible de les moduler par ajout des connaissances supplémentaires, variables selon les utilisateurs

et comme connaissances additionnelles les degrés de prototypicalité entre deux entités cognitives, c'est-à-dire des degrés de représentativité d'une entité par rapport à l'autre.

3.6. Enrichissement sémantique des requêtes dans les moteurs de recherche

Cette technique de reformulation permet d'exploiter la richesse des phrases définitoires pour augmenter la pertinence du résultat retourné (Rachdi *et al*., 2011). Elle permet de reformuler les requêtes composées d'un seul mot. Elle cherche sur Internet les phrases de définition de ce mot, après un module permet d'extraire les mots qui accompagnent mot dans les phrases de définition, appelés des mots associés. Une fois trouvés, un autre module effectue un traitement sur ces mots pour choisir ceux qui sont proches sémantiquement. C'est ces derniers qui seront utilisés pour enrichir la requête initiale.

4. Notre approche

La requête utilisateur n'est pas toujours bien formulée, donc il faut prévenir une reformulation pour augmenter la pertinence, augmenter la précision et diminuer le bruit. Pour arriver à cet objectif plusieurs techniques existent déjà, l'enrichissement qu'on propose entre dans le même cadre en exploitant la richesse des phrases de définition pour récupérer des mots communs qui représentent le domaine d'intersection entre les mots de la requête initiale.

Définition

Pour chaque mot de la requête initiale, on trouve un ensemble de phrases de définitions et chacune d'elles contient une liste de mots. Un mot commun c'est l'intersection entre les différentes listes.

Soit donc un ensemble de mots représentant la requête initiale. Quels sont les mots communs entre ces mots ? Comment peut-on les trouver et sur quel critère peut-on se baser pour les filtrer ?

Cet article essayera de répondre à ces questions en se basant sur les phrases de définition.

Dans un contexte général, une phrase de définition est porteuse de plusieurs informations qui donnent plus de détails et de précisions sur le mot définit. De ce fait elle se compose du mot définit et d'un ensemble de mots en relation appartenant généralement au même domaine que le mot définit. Le travail présenté dans (Rachdi *et al*., 2011) utilise ce type de phrases pour enrichir la requête utilisateur en utilisant les mots dits « proches ». Malgré que la méthode a donné de bons résultats elle contient quelques lacunes qu'on a essayé de prendre en considération dans cette nouvelle approche.

Parmi ces lacunes :

- L'approche permet de reformuler des requêtes utilisateurs composées d'un seul mot, alors que réellement elles se composent d'un ou plusieurs mots.
- Elle se base sur la fréquence d'apparition des mots et les mots sélectionnés sont ceux qui se répètent souvent dans la définition. Cette une mesure est relative et non pas absolue car dans un document on peut trouver des mots qui se répètent souvent mais qui apportent peu d'information.

- Pour la variable moyenne utilisée pour sélectionner les mots proches, elle est insuffisante car elle pourra être influencée par les valeurs aberrantes.

Donc la nouvelle approche permet de reformuler des requêtes utilisateurs composées de 1 à n mots en utilisant une autre technique pour sélectionner les mots utilisés pour la reformulation.

On commence dans cette approche par la recherche des phrases de définition de chaque mot contenu dans la requête initiale. A partir des phrases obtenues, on sélectionne les mots associés pour chacun des mots. On obtient alors pour chacun de ces derniers un ensemble de mots associés, puis on cherche l'intersection entre les ensembles des mots associés. Les éléments communs seront utilisés pour l'enrichissement de la requête.

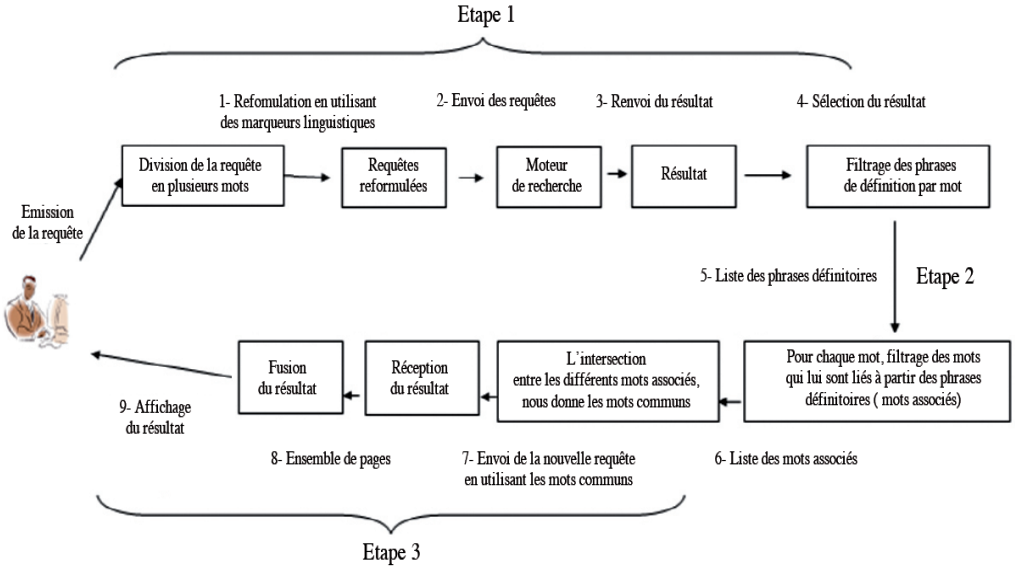


Figure 1 : Processus général de l'approche

En détaillant ce processus, on se retrouve avec les étapes suivantes :

4.1. Première étape : extraction des mots de la requête initiale et recherche des phrases de définition

La question posée à ce niveau c'est comment extraire les documents contenants des phrases de définition à partir d'Internet ?

Dans cette étape, au lieu de lancer la requête telle qu'elle est formulée par l'utilisateur, on cherche les formes possibles des phrases de la définition des mots recherchés.

Pour le faire, au moment de la réception de la requête, on la divise en plusieurs mots puis on procède à une reformulation par réinjection pour chacun des mots. Ce type de reformulation consiste à utiliser des marqueurs utilisés dans (Véronique *et al*., 2004) pour le repérage des énoncés définitoires, Quatre types de marqueurs ont été distingués, définissant quatre groupes de patrons :

- Les marqueurs métalinguistiques à utiliser indépendamment (au nombre de 9) : appeler, baptiser, définir comme, dénommer, dénoter, désigner, nommer, signifier, vouloir dire.
- Les marqueurs métalinguistiques nominaux (11) : appellation, acception, concept, dénomination, désignation, expression, mot, nom, notion, terme, vocable ; à associer à un verbe support parmi : appliquer, donner, employer, prendre, porter, recevoir, référer, renvoyer, réserver, utiliser.
- Les marqueurs lexicaux n'étant pas explicitement métalinguistiques, ou ceux de reformulation (21) : c'est-à-dire, en d'autres termes, soit, à savoir, en quelques sortes, une sorte de, enfin, il s'agit de, entendre par, vouloir dire, indiquer, comme, dit, par exemple, autrement dit, même chose que, équivaloir à, employer pour, marque, expliquer, préciser.
- Les ponctuations : parenthèses, guillemets et tirets d'incise sont également mentionnés dans la littérature. Vu que des paradigmes (aide à la modélisation horizontale de l'ontologie) et des hyperonymes peuvent aussi être extraits, ces quatre schémas ont été inclus à l'évaluation de la méthode même s'ils n'étaient pas exclusivement ciblés sur l'extraction de définition.

Dans notre cas on a utilisé les trois premiers marqueurs pour reformuler la requête initiale. La requête reçue va avoir donc une reformulation initiale pour chaque mot. On obtient alors plusieurs requêtes reformulées qui seront envoyées au moteur de recherche, le résultat envoyé sera récupéré sous forme d'un ensemble de documents.

Soit la requête initiale :

$Q= \{ M_1, M_2, ..., M_n \}$

où M_i est le mot i de la requête.

On utilisant la reformulation par réinjection de pertinence, on obtient un ensemble de requêtes où chaque requête se compose d'un ensemble de formes de phrases de définition pour un mot

$Q_0= \{ Q_1, \ldots, Q_p / Q_p$ est la requête p envoyée$\}$

avec Q_i est la requête i envoyée.

Comme résultat de cette étape, on obtient un ensemble Ω_i d'éléments pour chaque mot i de la requête tel que :

$\Omega_i = \{d_1, ..., d_n\}$

avec n est le nombre de documents trouvés, et

d_j est le document j trouvé.

4.2. Deuxième étape : extraction des mots associés

Chaque document dj dans un ensemble Ω_i trouvé contient un ensemble de phrases, L'étape suivante consiste alors à extraire que les phrases de définition, le reste du document sera ignoré pour le moment.

Soit $\sum_k$ l'ensemble des phrases de définition pour un mot

$\sum_k = \{pd_1, \ldots\ldots, pd_m\}$

où m est le nombre total des phrases trouvées, et

Pd_m est la phrase m de l'ensemble $\sum_k$.

Maintenant qu'on dispose des phrases définitoires des mots, on va extraire les noms qui apparaissent dans la définition des mots recherchés. A la fin de cette étape, on obtient donc la liste des mots de la requête initiale avec une liste des mots qui apparaissent en définition de ces mots, ce qu'on appelle les mots associés.

Soit donc X_i, l'ensemble des mots associés pour chaque mot de la requête

$X_i = \{ Ma_i \quad 0 < i < L \}$

avec L est le nombre de mots trouvés, et

Ma_i : le mot associé i.

Exemple :

pollution									
Environnement	polluant	Loi	Gaz	Cœur	Monde	Source	voiture	……	……
santé									
Vie	Développement	Secteur	Population	projet	Economie	Produit	Aspect	……	……
Clavier									
Avenir	organe	information	touche	invité	matière	moyen	personne	……	……

Figure 2 : Extrait de la liste des mots associés trouvés pour les mots pollution santé et clavier

L'objectif de cette étape est de savoir pour chaque mot appartenant à la requête initiale, la liste des mots qui lui sont associés, ce sont des mots qui apparaissent dans leurs phrases de définition et donc qui les représentent le plus.

4.3. Troisième étape : recherche de mots communs

Dans cette étape on dispose alors d'un ensemble de mots associés pour chaque mot, on cherche l'intersection entre les différentes listes des mots associés.

Soit donc MC la liste des mots communs trouvés :

$$\mathbf{M}_C = \bigcap_{i=0}^{n} X_i$$

X_i est la liste des mots associés pour le mot i.

Exemple des mots communs trouvés pour les mots pollution et santé :

N est le nombre de mots de la requête initiale.

Pollution et santé : environnement, source, monde, habitants, organisme, compagnie, nuisance, eau, marché, facteur, vie, consommation, ….

Finalement, on exploite la liste des mots communs trouvés pour l'enrichissement. On envoi donc plusieurs requêtes et chacune se compose des mots de Q_0 avec un mot commun trouvé.

Algorithme du processus de la recherche des mots communs

Entrées : un ensemble de document

Sortie : une liste des mots communs

```
Pour i allant de 0 jusqu'a Nbdocuments faire {
  Tant que  ( != fin de fichier) {
            X <—  extrairephdef ( ) ;
            Tant que (x non vide)  {
                      M(j) <—  extrairemot (x) ;
                      J++ ;
            }
  }
}
```

M_n est la liste des mots proches pour le mot n, et

M_C est la liste des mots communs.

On suppose que la requête se compose de deux mots

L=0

```
Pour i allant de 0 jusqu'a  M1.Length faire {
  Pour j allant de  0 jusqu'a M2.length faire {
            Si   M1(i) = M2(j) {
                      MC(L) = M1(i)
                      L = L+1
            }
  }
}
```

5. Résultat

Comme présenté dans les sections précédentes, cette technique de reformulation permet d'enrichir les requêtes en utilisant un type spécifique des phrases issues d'Internet c'est les phrases définitoires ou bien phrases de définition. Ces phrases contiennent des mots liés par différents types de relations (horizontales, verticales, …). Le tableau qui suit montre un exemple d'une requête utilisateur composée de mots clavier et santé (voir tableau 1) avec les mots communs trouvés.

Mots	Mots associés	Mots communs
Santé	Etat, physique, mental, social, infirmité, domaine, environnement , source, facteur, travail, terme, garantie, organisme, sécurité, frais, organe, manière, ville, eau, crise, événement, instrument, homme, pays, point, ensemble, arme, affaire, gouvernance, critère, beauté, charge, virale, projet, gain, main, cheval, table, stratégie, impôt, montant, payeur, luxe, médecin, monde, processeur, population, moyen, ressource, but, consommation, médicale, chantier, budget, année, responsable, politique, programme, pratique, système, logement, qualité, développement, offre, distributeur, approche, alimentation, air, temps, marché, thème, compagne, vie, étapes, naissance, enfant, mariage, richesse, secteur, industrie, profit, capitale, foi, processus, contrôle, groupe, individu, besoin, milieu, progrès, économique, aspect, remboursement, soins, matière, âge, examen, personne, centre, réseau, régional, innovation, produit, médecine, valeur, succès, guerre, mondiale, droit, nourriture, maladie, habitant, riche, voiture, bonheur	Etat, organe, système, point, travail, main, matière, moyen, marché,
clavier	organe, touche, information, ordinateur, périphérique informatique, nom, commune, province, département, personnalité, réalisateur, acteur, film, zone, souris, invité, pc, US, word, notepad, logiciel, azerty, anglais, avenir, tablette, aide, lumière, ami, fou, dingue, système, inconvénient, international, jeu de caractère, langue, lettre, qwerty, type, test¸ bon, mode, conférence, philosophie, périphérique d'entrée, écran, périphérique de sortie, pression, interruption, usage, disposition, version, semaine, point, câble, méthode, travail, cause, infection, fait, bureau, reste alimentaire, bactérie, animal, période, état, plastique, frappe, main, position, raison, utilisation, afficheur, maitre, école, raccourci, fichier, surface, matière, doigts, moyen, option, peau, marché, finition, ensemble robuste, rage, windows, dispositif, action, direction, précision, interface, communication, homme-machine, affichage, position, focus, utilisateur, choix, côté, avis public, français, cinéma, protecteur, boitier, capot, personne, machine, accessoire	

Tableau 1 : Liste des mots communs trouvés entre les mots clavier et santé

On utilisant cette technique de reformulation, le résultat retourné montre un taux de pertinence élevé par rapport à l'envoi de la requête sans reformulation, l'analyse de 30 premières pages retournées par notre approche a montré qu'un document parmi les 30 est non pertinent et pour le reste ce sont des documents qui peuvent constituer le résultat souhaité par l'utilisateur. L'approche proposée semble donc intéressante car elle permet de diriger l'utilisateur et ce cibler le résultat.

Exemple en utilisant les mots santé et clavier :

- 9 sur 20 non pertinents sans reformulation.
- 2 sur 20 en utilisant notre approche de reformulation.

Le lien suivant est pertinent et semble le lien cherché par l'utilisateur et ne s'affiche pas dans les dix premières pages de google sans reformulation www.motioncomputing.fr/resources/MWkeyboard_FR.pdf.

6. Conclusion et perspectives

Le travail présenté dans cet article s'inscrit dans le cadre de développement d'un outil d'aide a l'enrichissement des requêtes par phrases de définition dans un système de recherche d'information. Nous avons utilisés des marqueurs pour repérer les phrases de définition puis on a utilisé les mots communs pour la reformulation.

Le but de ce traitement est d'améliorer la précision des requêtes reçus par les SRI. La faisabilité de la méthode a été testée en utilisant le moteur de recherche Google.

Nous avons montré que l'utilisation des mots communs est une vraie solution pour enrichir les requêtes. L'approche proposée permet de borner la requête et cibler le résultat.

- Malgré que le résultat obtenu semble important, mais il reste un travail important dans la phase finale qui consiste a fusionner et trier le résultat, dans cette approche on utilise l'algorithme de google mais on travail sur un nouvel algorithme pour fusionner le résultat.
- Les mots communs sont utilisés dans notre cas pour enrichir la requête, on peut aussi exploiter les relations entre les mots de la requête initiale pour définir le domaine de la requête et donc pour l'enrichissement.

Référence

M. Boughanem, C. Chrisment, C. Soule-Dupuy (1999). Query modification based on relevance backpropagation in adhoc environment. Information Processing and Management, vol. 35, pp. 121-139.

I. Ruthven, M. Lalmas (2003). A survey on the use of relevance feedback for information access systems. Knowl. Eng. Rev., 18(2):95-145.

Y. El Younoussi, A. Doukkali Sdigui, E. Ben Lahmer (2010). Promoting the relevance of the research information systems via a query reformulation process. *The 6th International Computing Conference in Arabic (ICCA'10)*, mai 2010, Tunes, Tunisie.

E. Ben Lahmer, A. Doukkali Sdigui, Y. El Younoussi (2010). The research of terms definitions by metasearch. *The 6th International Computing Conference in Arabic (ICCA'10)*, mai 2010, Tunes, Tunisie.

X. Aimé, F. Fürst, P. Kuntz, F. Trichet (2010). Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées. Atelier de la Personnalisation du Web, *10ème journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010)*.

M. Rachdi, H. Ben Lahmer, H. Labriji (2011). Semantic enrichment of queries in search engines. Extraction et Gestion des Connaissances (EGC-M 2011), novembre 2011, Tanger, Maroc.

V. Malaisé, P. Zweigenbaum, B. Bachimont (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *Traitement automatique des langues naturelles (TALN 2004)*, Fès, Maroc, avril 2004, pp. 269-278.

TICAM'12, p. 225-232

Matching des Documents XML par la Mesure de Similarité à base Wordnet

Fatiha Djahafi **Abdelkader Haouas**
Université des Sciences et de la Technologie d'Oran USTO-MB Algérie
fatiha.or@hotmail.com haouasab@yahoo.fr

Résumé

De nos jours, XML est devenu un standard incontournable pour la représentation et l'échange de données sur le Web. De plus, non seulement les collections de documents XML sont réutilisées, mais le volume de leurs échanges s'accroît continuellement. Le problème engendré par ce type de document est lié à la nature de leur contenu. Nous nous intéressons plus particulièrement au matching, qui est nécessaire pour la collection de documents xml. Le matching est par définition un processus qui vise à identifier et découvrir les correspondances sémantiques entre différents formats de documents tels que les documents xml. Notre objectif consiste à être en mesure de réaliser un matching entre les termes des documents xml en utilisant le thesaurus WordNet et en s'appuyant sur les différentes techniques de mesure de similarité à base de Wordnet.

1. Introduction

Le monde informatique compte beaucoup de données aux formats très hétérogènes, autrement dit utilisant des modèles différents pour la représentation de l'information. Donc il est nécessaire d'établir une méthode de correspondance ou d'appariement entre ces données. XML est devenu un standard pour la représentation et l'échange de données. Le nombre de documents XML échangés augmente de plus en plus, et la quantité d'information accessible aujourd'hui est telle que les outils, même sophistiqués, utilisés pour rechercher l'information dans les documents ne suffisent plus. D'autres outils permettant de synthétiser ou classer de larges collections de documents sont devenus indispensables.

De nos jours, plusieurs approches gèrent la problématique de l'hétérogénéité entre les documents xml et de leur intégration par des méthodes de matching (Sauvagnat, 2004). Parmi celle-ci, plusieurs méthodes emploient un modèle de similarité sémantique pour rétablir les relations entre les concepts des documents, c'est-à-dire pour effectuer un matching entre ces documents. Dans ce contexte, le matching des documents xml s'agit d'une technique qui effectue la découverte de correspondances sémantiques entre les éléments et les attributs des documents pour résoudre ce problème d'hétérogénéité. Par ailleurs, nous proposons d'utiliser la ressource sémantique wordnet qui nous permettra d'identifier et mesurer les éléments similaires sémantiquement et de les prendre en considération lors de la phase de matching. En effet, la similarité joue un rôle crucial dans nombreux domaines de recherche. La Similarité sert comme un principe d'organisation par lequel les individus classent des objets, des concepts de forme, et fassent des généralisations.

Nous présentons dans cet article une approche de matching des documents xml basée sur un modèle de mesure de similarité wordnet pour découvrir les correspondances sémantiques entre les éléments des documents. L'objectif principal de notre travail est de proposer une approche de matching. Elle utilise une technique de mesure de similarité basée sur le thesaurus WordNet (Zerdazi et Lamolle, 2006). Cette approche comporte quatre phases fondamentales : Extraction de documents XML en utilisant un parseur, génération de l'arbre correspond à sa structure, la recherche de similarité par l'utilisation de WordNet et calcul de la mesure de similarité globale à fin de découvrir des matching.

Cet article est organisé de la manière suivante : outre l'introduction, la section 2 est un état de l'art qui présente les méthodes de matching et se focalise sur les travaux de mesure de similarité. Dans la section 3, nous présentons notre approche avec ses différentes étapes. La section 4 définit la mesure de similarité pour le matching. La section 5 est une étude expérimentale qui discute les résultats de notre méthode. Enfin, la section 6 conclut le travail effectué et présente les perspectives à venir.

2. Etat de l'art

Dans cette section, nous allons présenter les approches principales de matching et les différentes techniques de mesure de similarité.

2.1. Matching

Le matching est considéré comme une solution au problème d'hétérogénéité des documents semi-structurés, il permet de découvrir les correspondances entre les éléments de des différentes documents, il peut se faire manuellement mais avec la complexité et la taille de corpus cette manière devienne fastidieuse, automatiquement ou semi-automatiquement sont les meilleures façons de le faire.

Plusieurs travaux de matching ont été proposés dans la littérature (Zerdazi et Lamolle, 2006; Rahm et Bernstein, 2001; Shvaiko et Euzenat, 2005). En effet, dans (Shvaiko et Euzenat, 2005) les auteurs ont proposé une approche de classification de différentes techniques de Matching. Il existe un certain nombre d'algorithmes de matching qui cherchent à déterminer les correspondances entre les éléments de documents (Rahm et Bernstein, 2001). Ces algorithmes utilisent souvent des méthodes appliquées aux éléments des documents et des techniques structurelles. Nous retenons :

- **Techniques terminologiques** (Madhavan *et al*., 2001; Monge et Elkan, 1996): Elles sont souvent utilisées afin de déterminer le matching des noms et de leurs descriptions. Ces méthodes se basent sur la comparaison des termes ou des chaînes de caractères ou bien les textes.
- **Techniques linguistiques** : Ces méthodes exploitent essentiellement des propriétés expressives et productives de la langue naturelle (Maynard et Ananiadou, 1999). Les informations exploitées peuvent être celles intrinsèques (des propriétés linguistiques internes des termes telles que des propriétés morphologiques ou syntaxiques) ou celles extrinsèques (employant des ressources externes telles que des dictionnaires, des thésaurus).

- **Les méthodes structurelles internes** (Monge et Elkan, 1996) : elles calculent la similarité entre deux concepts en exploitant les informations relatives à leur structure interne (restrictions et cardinalités sur les attributs, valeurs des instances). Dans la plupart des cas, ce sont les informations concernant des attributs de l'entité, telles que la cardinalité des attributs, les caractéristiques des attributs ou les autres types de restriction sur les attributs.
- **Les méthodes structurelles externes** (Resnik, 1999; Wu et Palmer, 1994): exploitent les relations entre les entités elles-mêmes, qui sont souvent des relations de subsomption (« is-a ») ou de méréologie (part-whole). Avec ces relations, les entités sont considérées dans des hiérarchies et la similarité entre elles est déduite de l'analyse de leurs positions dans ces hiérarchies. L'idée de base est que : si deux entités sont similaires, leurs voisines pourraient également être d'une façon ou d'une autre également similaires.

2.2. Les techniques de mesure de similarité

La similarité joue un rôle très important, en particulier dans le processus de matching. Elle se rapporte à la comparaison des éléments des documents. Elle renvoie une valeur numérique indiquant si les deux éléments de document ont un degré élevé ou bas de similitude. La notion de similarité sémantique (ou son inverse : distance sémantique) est utilisée pour exprimer la ressemblance entre des concepts. Certaines mesures de similarité sémantique, ont été proposées en utilisant les ressources sémantiques disponibles.

Dans la littérature (Slimani *et al*., 2006), plusieurs travaux sur la mesure de similarité sémantique ont été développés dans différents contextes. On peut distinguer trois grandes familles d'approches de calcul de ces mesures :

- Les approches basées sur le comptage de lien (sur les arcs) : les auteurs Wup *et al*. (Wu et Palmer, 1994; Rada *et al*., 1989; Lee *et al*., 1993) sont basés sur l'hypothèse que plus il y a de liens entre deux concepts de documents plus ils sont similaires. Les auteurs comparent cinq mesures de similarités ou distances sémantiques utilisant WordNet (Fellbaum, 1998) (où la relation « is-a » est restreinte aux noms et verbes). Un état de l'art complet est présenté par Patwardham (Patwardham, 2003) où sont comparées ces différentes mesures entre elles par rapport à des évaluations faites par des sujets humains.
- Les approches basées sur les nSuds (Lee *et al*., 1993) utilisant typiquement des mesures du contenu informationnel (CI) pour déterminer la similarité conceptuelle. En plus, la similarité entre les concepts est déterminée par le degré de partage de l'information. Plus précisément le contenu informationnel se calcule de la manière par la formule suivante : $CI(c) = - \log(P(c))$
où P(c) est la probabilité de retrouver une instance du concept c. Ces probabilités sont calculées par : frequence(c)/N où N est le nombre total de concepts.
- L'approche hybride (Resnik, 1999; Leacock et Chodorow, 1998) qui combine entre les deux premières approches. Ces méthodes sont fondées sur un modèle mixte qui combine entre des approches basées sur le comptage des liens en plus du contenu

informatif qui est considéré comme facteur de décision. La mesure adoptant cette méthode est basée sur la combinaison d'une source de connaissance riche (thesaurus) avec une source de connaissance pauvre (corpus). La formule proposée par (Jiang et Conrath, 1997) est définie par l'inverse de la distance sémantique.

$$\text{sim}_{\text{Jiang_Corath}}(c_1, c_2) = \frac{1}{\text{distance}\ (c_1, c_2)}$$

Sachant que la distance entre c1 et c2 est calculée par la formule suivante :

$$\text{distance}\ (c_1, c_2) = 2 \times \text{CI}(psc(c_1, c_2)) - (\text{CI}(c_1) + \text{CI}(c_2))$$

Finalement on peut dire que la performance des mesures de similarité sémantique dépend de la qualité de la ressource sémantique et du corpus utilisé.

3. Notre Approche

Notre approche s'appuie sur une mesure de similarité à base de WordNet. Nous nous intéressons plus particulièrement au matching, qui consiste à trouver et découvrir les correspondances sémantiques entre différents éléments de documents XML. L'avantage de ces documents est qu'ils possèdent une structure qui facilite leur présentation, ainsi que leur interprétation et leur exploitation dans des contextes présentant différents besoins. Cette mise en correspondance permet de calculer la similarité entre les documents de la collection en se basant sur le WordNet.

Cette approche est constituée de quatre étapes. Dans la première étape, il s'agit d'extraire la structure donnée par les balises d'un parseur. La deuxième étape concerne la transformation en représentation arborescente. Dans la troisième étape, on se sert de WordNet dans un premier temps pour récupérer les différents sens possibles pour les unités sémantiques. La quatrième étape, on calcule les similarités entre les différents sens de ces unités en se basant sur les relations sémantiques et les pondérations des termes afin de réaliser le matching.

3.1. Utilisation de wordnet

WordNet est une ressource lexicale de la langue anglaise développé à l'Université de Princeton disponible en format électronique[1], il regroupe des termes (noms, verbes, adjectifs et adverbes) en ensembles de synonymes appelés synsets. Un synset regroupe tous les termes dénotant un concept donné. Les synsets sont reliés entre eux par des relations sémantiques. Pour le calcul de la similarité linguistique, la fonction Syn(c) calcule l'ensemble des Synsets de WordNet du concept c; soit $S=Syn(c_1) \cap Syn(c_2)$ l'ensemble des sens communs entre c_1 et c_2 à comparer, la cardinalité de S est :

$$\lambda\ (S) = |Syn(c_1) \cap Syn(c_2)|$$

1 http://wordnet.princton.edu/

Soit min (|Syn(c_1)|,|Syn(c_2)|) le minimum entre les cardinalité des deux ensembles Syn(c_1) et Syn(c_2), alors la similarité entre deux concepts c_1 et c_2 sera définie comme suit:

$$Sim\ (c_1, c_2) = \frac{\lambda\ (S)}{min\ (|Syn(c_1)|, |Syn(c_2)|)}$$

Cette mesure retourne 1.0 si au moins c_1 est le seul synonyme de c_2 ou c_2 est le seul synonyme de c_1.

4. Calcul de similarité pour le matching

Le calcul de la similarité se fait entre les ensembles de documents. Les documents sont représentés par des ensembles de vecteurs de termes. Chaque unité de contexte génère un vecteur. Les poids des termes W_{ij} (Robertson, 1997) sont calculés en fonction de leur distribution dans les documents par la formule suivante : $W_{ij}=TF*IDF_{ij}$. La similarité globale de matching est calculée de la manière suivante :

$$Sim_G\ (A_1, A_2) = \frac{\sum^{n}_{i=1}\ \sum^{m}_{j=1} Sim_L(t_{1i}, t_{2j})}{Max\ (\ |A_1|, |A_2|)}$$

où A_1 et A_2 représentent deux ensembles des termes de même catégorie. Les similarités locales des couples (t_{1i}, t_{2j}) soient déjà calculées tel que les termes t_{1i} et t_{2j} appartiennent respectivement aux documents A_1 et A_2. $|A_1|$, $|A_2|$ étant respectivement les tailles (nombres de termes n et m) des documents, la division par Max ($|A_1|$, $|A_2|$) permet de normaliser le résultat de la sommation. $|A_1|= n$ et $|A_2|= m$.

Cette fonction donne comme résultat les couples de l'ensemble D = A_1 x A_2. Les couples (t_{1i}, t_{2j}), intervenant dans le calcul, doivent présenter les meilleures mesures de similarité. Le résultat sera le matching des documents A_1 et A_2.

5. Expérimentation

On tente de comparer les deux articles d'un seul document par notre méthode matching :

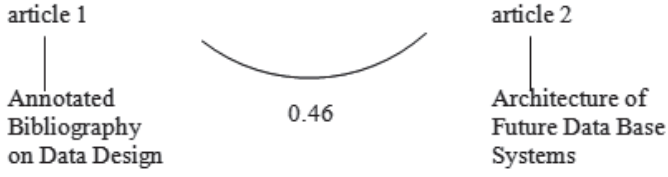


Figure 1 : Exemple de comparaison

Chaque article de document est composé des termes t_j. La pondération des termes permet d'exprimer le pouvoir discriminant d'un terme t_j dans le document d_j. Le pouvoir discriminant d'un terme est sa capacité à distinguer les documents les uns des autres. Ce calcul de poids s'inspire de la méthode TF-IDF qu on applique aux documents. Le tableau suivant montre exemple de la force de discrimination de chaque terme :

Termes	W_{ij}
t_1	0.33
t_2	0.47
t_3	0.38
t_4	0.47
....	....

Tableau 1 : Les occurrences des termes

Ensuite, nous avons implémenté la similarité de description entre ces termes par l'intégration de wordnet. Cette similarité basée sur les relations sémantiques entre les concepts. Les valeurs de similarités entre les termes des 2 documents sont données dans la table suivante :

		Article 1				
		architecture	*future*	*data*	*base*	*systems*
Article 2	*annotated*	*0*	*0*	*0*	*0*	*0*
	bibliography	*0*	*0.09*	*0*	*0*	*0*
	data	*0*	*0*	*1*	*0*	*0.25*
	design	*0*	*0*	*0*	*0*	*0*

Tableau 2 : Table de similarité dans le wordnet

Nous pouvons remarquer à partir de ce point de résultats que la similarité wordnet entre deux concepts similaires est égale à 1, par exemple (data, data) = 1. Nous constatons que cette mesure de similarité donne de bons résultats. Par exemple, la similarité entre (data, data) de deux concepts similaires syntaxiquement et sémantiquement est égale à 1, alors que pour (*Bibliography*, *Future*) qui sont deux concepts syntaxiquement et sémantiquement éloignés l'un de l'autre, notre approche nous a retournée une valeur très faible de 0.09.

Nous avons effectué un test, on a pris par exemple 4 articles de document du corpus ACM Sigmod et on leur a appliqué notre mesure pour réaliser le matching correspond. Le tableau suivant récapitule les résultats de premier test obtenu par notre algorithme de matching basé sur le wordnet.

Doc1	Doc 2	Matching
A_1	A_2	0.46
A_1	A_3	0.13
A_1	A_4	0.35
A_2	A_3	0.18

Tableau 3 : Résultats de nos tests

Les résultats obtenus par cette méthode conforte bien notre idée de base et consolide fortement notre contribution. Par exemple, nous constatons que le matching $(A_1, A_3) = 0.13$, et la similarité $(A_2, A_3) = 0.18$. On remarque que A_1et A_2 sont plus proche sémantiquement par rapport A_1 et A_3. Donc les articles publiés dans ce corpus sont de mêmes catégories.

6. Conclusion

Le but à atteindre dans ce présent travail est de mesurer les similarités entre les unités sémantiques de documents XML afin de déduire que cette unité a été découverte dans un autre document de même corpus avec un certain degré de similarité.

Dans la suite de ce travail, nous envisageons de prendre en compte des documents xml volumineux afin d'effectuer les matching les plus complexe des corpus et de couvrir plus de cas de correspondance existants. Il serait intéressant aussi d'utiliser d'autres ressources que le wordnet pour améliorer les relations sémantiques. La validation de notre approche fera aussi parti de notre futur travail.

Référence

Fellbaum C. (1998). *WORDNET. An Electronic Lexical Database*. The MIT Press.

Jiang J., Conrath D. W. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*. International Conference Research on Computational Linguistics (ROCLING X).

Leacock C., Chodorow M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*. WordNet: An Electronic Lexical Database, C. Fellbaum, MIT Press.

Lee J.H., Kim M.H., Lee Y.J. (1993). *Information Retrieval Based on Conceptual Distance in IS-A Hierarchy*. Journal of Documentation 49, pp. 188-207.

Madhavan J., Bernstein P., Rahm, E. (2001). Generic schema matching with cupid. The 27th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc, pp. 49-58.

Maynard D. G., Ananiadou S. (1999). Term extraction using a similarity based approach. Recent Advances in Computational Terminology. John Benjamins.

Monge A., Elkan C. (1996). The field-matching problem: algorithm and applications. The 2nd International Conference on Knowledge Discovery and Data Mining, pp. 267-270.

Patwardham S. (2003). Incorporating Dictionary and Corpus Information in a Measure of Semantic Relatedness. M.S. Thesis.

Rada R., Mili H., Bichnell E., Blettner M. (1989). Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, pp. 17-30.

Rahm E., Bernstein P. (2001). A survey of approaches to automatic schema matching. VLDB Journal, 10(4): 334-350.

Resnik P. (1999). Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, n°11, pp. 95-130.

Robertson S.E. (1997). *The probability ranking principle in IR*. Journal of Documentation, 33 (4): 294-304.

Sauvagnat K. (2004). *XFIRM : Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés*. CORIA' 04, Toulouse, France, pp.121-142.

Shvaiko P., Euzenat J. (2005). A Survey of Schema-based Matching approaches. Journal on Data Semantics IV, vol.3730, pp. 146-171.

Slimani T., Yaghlane B., Mellouli K. (2006). A new similarity measure based on edge counting. World academy of science, engineering and technology, Décembre 2006, vol. 17.

Wu Z., Palmer M. (1994). Verb semantics and lexical selection. The 32nd Annual Meeting of the Associations for Computational Linguistics, pp.133-138.

Zerdazi A., Lamolle M. (2006). *Intégration de sources hétérogènes par matching semi-automatique de schémas XML étendus*. 24ème Congrès informatique des organisations et systèmes d'information et de décision, Inforsid'06, Hammamet, Tunisie, pp. 991-1006.

TICAM'12, p. 233-235

Cours de Tamaziγt par Internet : Problèmes et Propositions

Saïd Chemakh Malika Sabri
Département de Langue et Culture Amazighes,
Université M. Mammeri de Tizi-Ouzou.
{chemakh_s, sabrim6}@yahoo.fr

1. Introduction

Nous proposons un projet de création d'un cours de Tamaziγt par Internet, pour répondre à des besoins réels et immédiats, certes plusieurs sites proposent des cours de Tamaziγt mais la majorité d'entre eux ne correspondent pas aux attentes des demandeurs.

Les demandes qui se font sur Internet viennent de divers horizons, hors les offres des providers d'Internet, restent limités et pour certaines d'entre elles sont nulles.

Pendant l'année (2007-2008), nous étions confrontés à l'enseignement de Tamaziγt à des étudiants titulaires d'un BAC mais arabophones natifs (essentiellement de la région de Tiaret).

Il fallait trouver des solutions pour remédier à cette situation : à savoir la méconnaissance de la langue berbère.

Un cours spécial a été crée pour y remédier.

Diverses méthodes ont étés testées : à savoir en premier la méthode « Tizi b uccen » : celle-ci s'avérant désuète est dépassée, les enseignants ont recours aux polycopies, élaborés par H. Oubagha pour ses cours pour arabophones, à la bibliothèque nationale. Mais il y a lieu de remarquer que seuls cinq (05) inscrits sur 27 ont pu tenir jusqu'à la 4ème année, et avoir leur licence.

Certes, si une bonne volonté anime ces personnes pour tenir le coup.

L'idée nous est venue de recourir en tic e-Learning comme apport et support d'aide à cette population estudiantine. Vue que Tamaziγt étant une discipline enseignée dans trois (03) universités et bénéficiant d'un statut de langue nationale, est devenue une filière comparable aux autres.

Avant de faire nos propositions sur un e-Learning de Tamaziγt, dressant d'abord un bilan de la présence de Tamaziγt, et de son enseignement sur le Web.

2. Présence des langues sur le Net

Dès l'avènement de l'Internet, la domination de la langue anglaise y est apparaît. Cette situation réside du fait que les centres de propagation du Net, restent les USA et le monde anglo-saxon, et que la formation des autres chercheurs de part le monde est faite par le biais de la langue anglaise. Ce n'est que dans les années (90) que d'autres langues telles que l'Espagne, l'allemand, le russe, etc., commencent à être plus utilisés en informatique, et par la même sur le Net.

De nombreuses langues se sont vues de fait exclues de cette NTIC pour diverses raisons :

- Langues essentiellement orales, sans traditions écrites attestées.
- Inconvénients des graphies utilisées avec les systèmes en usage sur le Net.
- Manques de moyens pour l'accès pour ces langues en NTIC.

Tamaziγt en fait partie de ses langues qui ont souffert du manque d'accès au Net, dès son départ. Deux raisons principales y président à cette situation :

1- Les caractères utilisés différent entre les diverses instances activant sur le terrain.

   Certaines d'entre elles désirent utiliser le Tifinaγ (néotifinagh) comme critère d'authenticité, d'autres préfèrent le caractère latin jugé « idéologiquement » plus occidental et donc vecteur de la modernité.

2- Les élites formées dans les domaines de l'informatique et de l'Internet (restent limitées).

Hormis quelques universitaires maîtrisant les rouages des disciplines telles que l'informatique et le Net peut ayant acquis des formations en langues et cultures amazighes ; ce que nous montrerons dans le paragraphe suivant.

3. Présence de Tamaziγt sur le Net

Avant l'apparition des sites web (www.kabyle.com) le Tamaziγt est déjà utilisé comme moyen de communication entre, essentiellement, les militants des cette cause.

Dès 1994, un consensus a était trouvé : à savoir recourir aux systèmes de la transcription de *BOULIFA*, ce système usité de graphie : « th pour ḍ », « d' pour d » a était en usage jusqu'aux années 60 lorsque *MAMMERI* ayant décidé de recourir aux systèmes des points souscrits sur les cinq (05) emphases et le ḥ, ainsi que l'insertion de chevron sur les affriquées : j et ch (ǧ, č), et l'insertion du ɛ et du γ.

Puisque plusieurs articles renvoyant à des cours de Tamaziγt sur les moteurs de recherches, peu d'entre eux en sont crédibles.

4. Les cours de Tamaziγt disponibles

Sur les moteurs de recherche comme: www.google.com et www.yahoo.com, des milliers de pages s'affichent, la déception est de voir à cette entrée du contenu vide ou redirigé vers d'autres sites publicitaires, etc.

Même les sites les plus consultés par des internautes amaziγ osent à peine proposer des cours de tamaziγt bien qu'ils y militent. Peu s'ils y sont investis dans l'enseignement/apprentissage de la langue.

5. Avantages et Inconvénients des sites d'enseignement de Tamaziγt

Les avantages de l'existence des cours de Tamaziγt sur le Net est cette présence même sur le web; donc accessible à tout internaute désireux de s'informer en/ou sur la langue berbère.

Les inconvénients résident dans la formation amateur, des webmasters qui ne réfléchissent pas à l'ensemble des problèmes que pose l'apprentissage de Tamaziγt.

L'apprentissage d'une langue maternelle (code oral/code écrit) diffère de l'apprentissage d'une langue second.

Qu'en est-il de Tamaziγt? Dois-t-on utiliser les mêmes méthodes pour les locuteurs natifs et les non-natifs?

L'aberration entretenue dans les programmes et manuels scolaires semble être entretenue comme les sujets du bac.

6. Propositions pour un e-learning de Tamaziγt

Enseigner une langue c'est aussi enseigner une culture, avec les moyens offerts par les technologies modernes, tout est possible: tout savoir est accessible à l'import quel être humain pour peu qu'il y ait accès en NTIC. Les personnes désireuses d'apprendre le Tamaziγt qu'ils soient locuteurs (code oral) ou non locuteurs doivent avoir les moyens d'y accéder aux perfectionnement ou à l'apprentissage de cette langue.

Nos propositions sont donc les suivantes:

La création d'un site avec accès en trois langues (Arabe-Tamaziγt-Français) :

- Tamaziγt: il s'agit essentiellement des cours destinés au perfectionnement des acquis scolaires (Unicode y est compris).
- Français: métalangage s'agissant des cours destinés soit à la préparation du BAC pro-français et/ou BAC lettre. Et sujets BAC ou BEM algériens.
- pour ce qui est de l'arabe, il s'agirai d'une création de cours interactifs (audio-audiovisuels,...) pour favoriser l'apprentissage de Tamaziγt essentiellement chez les étudiants arabophones orientés au département de langues et cultures amazighes.

L'Usage de la Langue Amazighe dans les Médias Algériens

Ouerdya Kireche
Université Mouloud MAMMERI de Tizi Ouzou, Algérie
kirkatia@yahoo.fr

Résumé

Nous proposons d'orienter notre contribution, que nous intitulons « L'usage de la langue amazighe dans les médias algériens », sur quatre axes principaux :

- La place de la langue amazighe dans le champ médiatique algérien;
- La qualité la langue amazighe utilisée dans les médias algériens;
- Attitudes du public à l'égard des programmes présentés en berbère;
- La diversité linguistique et culturelle est elle une chance ou un handicape pour la promotion de la langue amazighe à travers les médias algériens.

Introduction

Les moyens de communication, notamment ce qu'on appelle les médias lourds, sont d'une extrême impotence dans toutes les sociétés, et plus particulièrement dans les sociétés à tradition orale tel est le cas pour la communauté berbère d'une façon générale et la communauté kabyle[1] d'une façon particulière.

La communauté kabyle est caractérisée, pour longtemps, par une communication de type traditionnelle orale où l'échange d'informations et de communications se faisait de bouche à oreille.

Avec l'apparition de la radio nationale d'expression amazighe « la radio algérienne chaîne II », la population s'est vite attachée à cette radio qui reflète selon eux leur identité. Elle constitue une référence, c'est ce qu'on peut comprendre par leurs dires. Ex : Akka i d- nnan deg rradyu « c'est ce qu'ils ont dit à la radio ». Donc quand ils (les animateurs ou journalistes) disent quelque chose, il est indiscutable.

Cette radio, donc, constituait un moyen essentiel de l'action socioculturelle, économique et même politique. Les animateurs de cette radio avaient de multiples tâches : éduquer, cultiver, informer, ...

1 Le kabyle est le dialecte parlé en Kabylie. La Kabylie est « d'une superficie relativement limitée mais très densément peuplée, elle compte à elle seule les deux tiers des berbérophones algériens ».

Aujourd'hui, à côté du téléviseur, la radio envahie tous les espaces, surtout après sont intégration dans nos téléphones portables, résultat de l'évolution du champ médiatique.

Les systèmes de communication en Algérie connaissent, en fait, une évolution relativement importante.

Mais la question qui se pose est : Où est tamazight dans toute cette évolution ?

A côté de la radio algérienne chaîne II qui est, donc, la plus ancienne radio berbère en Algérie, l'état algérien a procédé, ces dernières années, à l'ouverture d'une télévision TV4 et de radios locales d'expression amazighe. Rappelant que la TV4 été inaugurée le 18 mars 2009, la radio Soummam, localisée à Béjaia, été inaugurée le 20 août 1996 (date importante dans l'histoire de la révolution algérienne), la radio de Bouira été inaugurée le 29 décembre 2008 (date qui correspond au premier moharram 1430), la radio de Tizi Ouzou, localisée été inaugurée le 01 novembre 2011(49ème anniversaire de la révolution algérienne) et la radio de Boumerdès qui a été inauguré le 05 juillet 2012 (50ème anniversaire de l'indépendance d'Algérie).

Ces médias qui sont sensés d'assurer un service public, et surtout de tenir compte de la réalité socioculturelle, économique et politique du public auquel ils s'adressent. Participer ainsi à la réhabilitation de la culture berbère. Cependant, selon nos informateurs, d'année en année, cette mission semblait loin d'être assurée, et l'auditeur ou le téléspectateur se désillusionna très vite, se rendant compte que les programmes de ces chaînes ne mettent pas en valeur toute la culture et la richesse du berbère. Les auditeurs et téléspectateurs ont du mal à s'identifier à ces chaînes.

Les images données des cultures des populations amazighes est souvent folkloriste et caricaturale et passéiste. Ce qui pousse les auditeurs ou les téléspectateurs à se plonger dans la haine de soi. Cette haine qui est nourrie aussi par l'aspect facultatif de l'enseignement de tamazight dans les collèges et lycées algériens.

La TV4, la seule télévision nationale d'expression berbère, et pour le but de renouer les liens et le contact entre tous les berbérophones d'Algérie, présente ces programmes en plusieurs dialectes à savoir le kabyle (dialecte répondu en Kabylie), le chaoui (dialecte répondu dans les Aurès), le targui (parlé dans le sud d'Algérie), le chenoui (parlé à Tipaza) et le mouzabit (parlé à Ghardaïa). Ce but n'est pas atteint, au contraire les téléspectateurs ont du mal à s'identifier à cette chaîne. Ils se perdent devant une mosaïque linguistique et culturelle qu'ils n'arrivent pas à comprendre facilement.

Certains de nos informateurs ont exprimé une distance à l'égard de certains animateurs et certains journalistes

Selon ces informateurs, ces présentateurs s'expriment dans une langue trop académique, savante et surtout chargée de néologisme, qu'ils ne comprennent pas forcement. D'ailleurs certains de nos informateurs le qualifient de berbère incompréhensif. Ces néologismes, sont le plus souvent utilisés d'une façon anarchique vue le statut de langue non normalisée du berbère. Jusqu'à ce jour il n'existe pas de glossaire de berbère normalisé.

En raison du manque de formation des animateurs et journalistes en matière de linguistique amazighe

Ces présentateurs se servent de leurs régolites et de leurs compétences linguistiques dans la présentation de leurs programmes ou journaux. Certains s'improvisent de traducteurs de textes souvent écrits en arabe ou en français. Traduction qui fait souvent défaut. Ex : tacriht n yilmezyen traduit de l'arabe carihat ccabab, sachant qu'on Kabyle tacriht signifie morceau de viande.

Sur le plan proprement linguistique

Nous trouvons que la langue de ces animateurs est teintée par le code switching, l'interférence et le calque.

Ex : Nhegga-d arupurtaj $ef lɛaddat n leqbayel « nous avons préparé un reportage sur les coutumes kabyles ».

Atas n yilemziyen i yeb$an ad d-yekriyyin des entreprises « beaucoup de jeunes qui veulent créer des entreprises ».

Nous avons remarqué un autre phénomène, certains animateurs, pour se distinguer ou pour se montrer intellectuels, ils font appel à d'autres langues, notamment au français.

Ex : Tettsellim $ef la famille is, ɛmumat-is, ses tantes. Tettmenni-as un joyeux anniversaire « elle passe le bonjour à sa famille, ses tantes paternelles, ses tantes (maternelles). Elle lui souhaite un joyeux anniversaire ».

Ce phénomène apparaît aussi dans les dénominations des émissions, qui devaient normalement être bien étudié.

Ex : Top ten, escale, mosaïque, ces émissions existaient des années avant mais avec des dénominations kabyles. On les dénommait respectivement, amyezwer gar tezlatin, amyezwer gar tezlatin, udem n taddart-iw. Les producteurs, pour moderniser ces émissions, ils ont procédé à les habiller avec des dénominations en français.

Conclusion

En conclusion, il faut dire que

- Tamazight est aujourd'hui dans l'obligation d'investir le domaine de communication.
- Les programmes présentés, doivent être réétudiés et réactualisés suivant les nouvelles réalités. Comme ils doivent refléter toutes les richesses linguistiques et culturelles de toutes les communautés auxquelles ils sont adressés.
- En attendant d'avoir un glossaire de berbère normalisé, les présentateurs doivent suivre une formation adéquate en linguistique amazighe. Comme ils doivent utiliser une langue populaire pour se rapprocher et attirer l'attention du récepteur.

Contribution à la Reconnaissance des Documents Tifinaghes

Mehdi Boutaounte Mohamed Fakir Belaid Bouikhalene
Equipe de Traitement de l'Information et Télécommunication, Département d'Informatique,
Facultés des Sciences et Techniques Béni Mellal, USMS, Maroc
boutaounte.mehdi@gmail.com
{fakfad, bbouikhalene}@yahoo.fr

Résumé

Récemment la reconnaissance des documents est devenue une nécessité primordiale pour plusieurs raisons : D'une part pour la sécurité des données existantes sur papier en vue de leurs durées de vie limitées, de plus le taux de destruction est élevé (insectes, feu, humidité). D'autre part pour économiser l'espace des archives. Dans ce travail, on présente une application basée sur les réseaux de neurones et implémentée sur des algorithmes dédiés à la reconnaissance des documents Tifinaghes.

1. Introduction

L'analyse et la reconnaissance d'images de documents englobent un ensemble de techniques informatiques avec comme but la reconstitution du contenu du document sous la forme de documents structurés. Les documents structurés couvrent deux catégories de documents : les documents imprimés et les documents manuscrits. Parmi les documents imprimés, nous distinguons les documents à structures simples et les documents à structures complexes. Dans cet article, nous nous intéressons aux documents à structures complexes.

La reconnaissance de documents s'applique à plusieurs langues écrites. La langue latine, Arabe, chinoise, etc ont reçu la plus grande attention de la part de chercheurs. En revanche, peu de travaux sur la reconnaissance de documents Amazighes ont été consacrés.

Cet article à pour sujet l'évolutivité des modèles dans un contexte interactif pour la reconnaissance de la structure physique et logique des documents Amazighes riches en structures et en variabilité.

En général la reconnaissance des documents est scindée en deux parties :

- Reconnaissance de la structure physique (André 2004) qui consiste à segmenter l'image de document en blocs (Figure 1).

Figure 1 : Structure Manhattan

Reconnaissance de la structure logique (André 2004) (Figure 2) : comprend l'usage des éléments d'intelligence artificielle afin de pouvoir classer les blocs extraits dans la partie précédente en utilisant les méthodes de classifications comme les méthodes statistiques, les réseaux de neurones, les réseaux bayésiens, les algorithmes génétiques, etc.

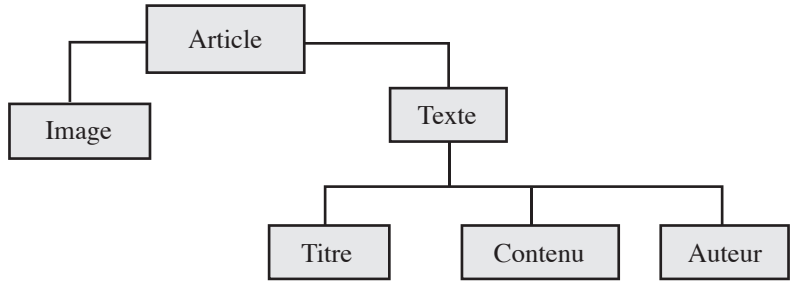


Figure 2 : Structure logique

Cet article est organisé comme suit : Dans la section 2, on présente les étapes de prétraitement qui englobe la réduction du bruit, la binarisation et la correction d'inclinaison. Dans la troisième section, on décrit la méthode de reconnaissance de la structure physique. L'extraction des mots et la reconnaissance de la structure logique sont décrites respectivement à la quatrième et cinquième sections.

2. Prétraitement

L'image acquise par un capteur est toujours accompagnée de parasites : bruits, inclinaison, etc. Le prétraitement qu'on a appliqué sur nos images de documents est divisé en trois grandes parties :

2.1. Réduction du bruit

Nous avons utilisé le filtre médian pour « nettoyer » les images en éliminant certains défauts parasites. L'opération consiste à que chaque pixel de l'image à filtrer est remplacé par la valeur médiane des pixels voisins.

2.2. Binarisation

La binarisation est une opération qui produit deux classes de pixels, représentées par des pixels noirs et des pixels blancs. La méthode choisie est celle adoptée par « OTSU » (Oanh, 2004) basée sur le calcul d'un seuil automatique en calculant l'histogramme donné par

$$h(i) = \frac{n_i}{\sum ni} \qquad (1)$$

où n_i représente le nombre de pixels de niveau i dans l'image.

La séparation se fait à partir de la moyenne et de la variance :

$$Moy = \sum_{i=1}^{k} i * h(i) \qquad (2)$$

$$\text{var} = \sum_{i=1}^{k} histogramme(i) \qquad (3)$$

Pour chaque valeur de k =1, …, 255, on effectue le calcul suivant :

On note mT= *moy* (256), où 256 est le nombre totale des niveaux de gris.

$$s^2(k) = \text{var}(k) * (1 - \text{var}(k)) * (mT * \text{var}(k) - moy(k))^2 \qquad (4)$$

Le niveau qui maximise la fonction critère (4) est considéré comme seuil pour la binarisation de l'image. Ainsi, la valeur de seuil est obtenue lorsque pour un k donné on a :

$$s^2(k) = \max(s^2(i)) \qquad (5)$$

2.3. Correction d'inclinaison

On applique l'algorithme de lissage RLS (Run Length Smearing) (Azokly, 1995) afin de détecter la forme rectangulaire du document. On applique le lissage par RLS sur l'image inclinée dans les quatre sens avant de grouper les résultats dans une seule image par « ET logique » (Figure 3).

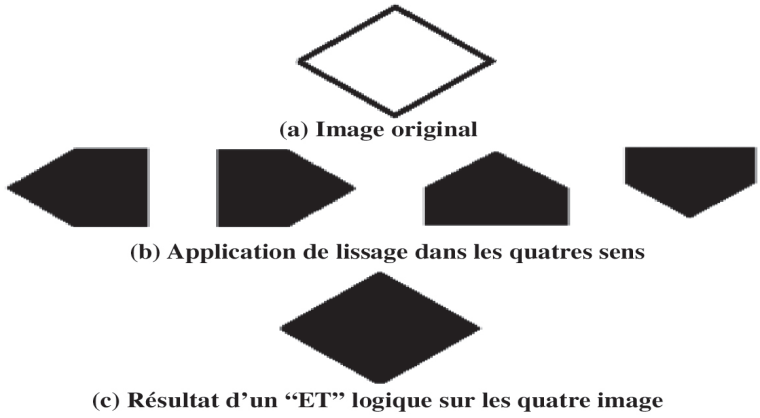


Figure 3 : Etapes d'application de lissage

Après le lissage on procède à la détection d'angle d'inclinaison comme suit :

- Détection des sommets du losange (Figure 4): Pour se faire, on s'est basé sur la valeur des pixels noires « 0 ». Donc on détecte le plus haut « 0 », le plus bas, le minimum à gauche et le maximum à droite dans la matrice.

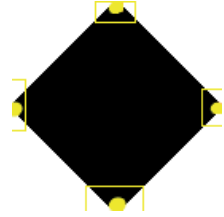


Figure 4 : Détection des sommets

- Détermination de l'angle d'inclinaison : Par la projection de deux points une sur l'axe horizontal et l'autre sur l'axe vertical. On obtient un troisième point (Figure 5) qui permit de déterminer « Θ » l'angle d'inclinaison.

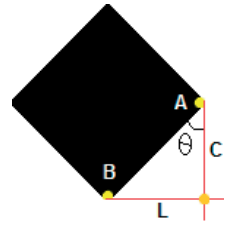


Figure 5 : Détermination de l'angle

Ce qui donne : $$\tan^{-1}\frac{L}{C} = \theta \qquad (6)$$

La figure 6 montre un document avant et après correction de l'inclinaison

Figure 6 : Correction de l'image

3. Reconnaissance de la structure physique

La reconnaissance de la structure physique des documents est la partie la plus compliquée, car elle présente une richesse et variance au niveau des composantes. Notre étude est consacrée à résoudre ce problème dans le cas des documents à structure complexe (les journaux, Magazines, etc.).

Notre algorithme est divisé en trois parties :

- Création des blocs (Figure 7) par usage d'algorithme de lissage RLS afin de supprimer, l'espace blanc entre les mots d'une ligne et les lignes d'un même bloc.

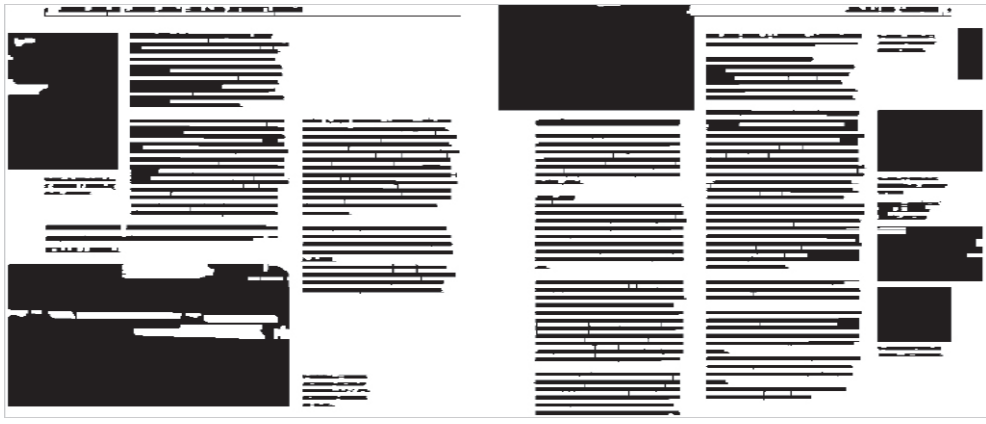


Figure 7 : Création de blocs

- Détection des blocs (Figure 8) : dans cette partie on utilise l'algorithme des composantes connexes (Caldairou, 2012) à 8-connections pour étiqueter chaque bloc avec un numéro unique.

Figure 8 : Détection de chaque bloc individuel

- Extraction des blocs (Figure 9) : on se base sur la matrice d'étiquetages de composantes connexes pour déterminer les coordonnés de chaque bloc afin de l'extraire.

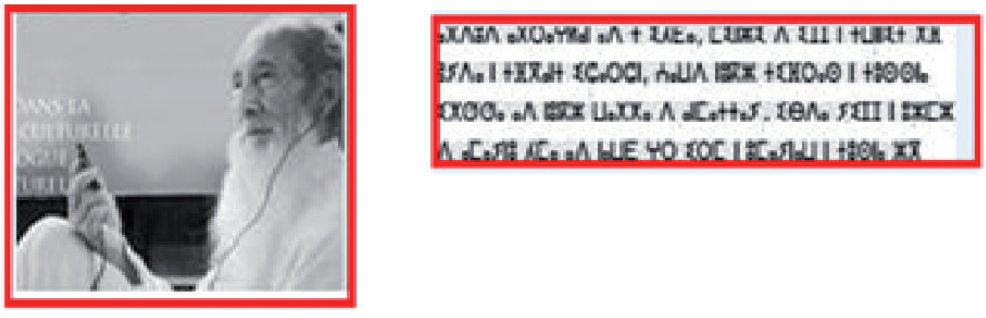


Figure 9 : Exemple de blocs détectés

Les résultats de segmentation sont bons dans le cas des articles simples, mais un problème se pose dans le cas des articles entourés par un cadre ou qui contient une image à l'intérieur du texte (Figure 10).

Figure 10 : Article cadré ou avec image

Pour remédier à ce problème, on a introduit une correction sur le fonctionnement de l'algorithme des composantes connexes.

On détermine les plus grandes parties dans chaque zone extraite après le premier traitement, puis on calcule la densité de cette partie par rapport à la taille totale de la zone.
La densité est représenté par :

$$Densité = \frac{C_x * C_y}{Z_y * Z_x} \qquad (7)$$

avec C_x, C_y sont les dimensions de la plus grande composante détectée, et Z_y et Z_x sont les dimensions du bloc.

Pour éviter les images qui ont une densité très élevée, on prend :

$$\text{Densité} < 0.9 \qquad (8)$$

Pour éviter les blocs de texte qui ont presque une égalité entre le nombre des lignes noires et blanches, on prend :

$$\text{Densité} > 0.5 \qquad (9)$$

Par la suite, on utilisera la même procédure pour extraire les composantes connexes de l'image, mais avant il faut changer l'étiquetage de telle manière à conserver uniquement la plus grande composante.

4. Extraction des mots

L'extraction des mots est la partie qui mène à une bonne classification afin de pouvoir utiliser une application de reconnaissance des caractères Tifinaghes pour extraire un document électronique complet.

La première phase de l'extraction consiste à segmenter l'image texte en lignes. Pour cela on a utilisé l'algorithme de RLS pour le groupement des mots de la même ligne, et détecté ce groupe de mots à l'aide de l'algorithme des composantes connexes et les extraire par la méthode de détection des rectangles déjà utilisée dans la section précédente.

Figure 11: Résultat de la segmentation des lignes

La suite consiste à extraire chaque ligne indépendamment afin de pouvoir détecter les mots, on utilise une projection (de la ligne) sur l'axe horizontal et on élimine les espaces inférieurs à un seuil déjà déterminé afin de ne pas diviser un seul mot en plusieurs parties ce qui donne le résultat de la figure 12. (Le taux de reconnaissance des mots calculer expérimentalement et varie entre 75% et 80%).

Figure 12 : Segmentation des mots

5. Reconnaissance de la structure logique

La reconnaissance de la structure logique des documents consiste à utiliser les outils d'intelligence artificielle pour donner à notre système les capacités de distinguer entre les blocs déterminés dans la patrie de reconnaissance de la structure physique.

La classification se fait avec un réseau de neurones (Todinca, 2008; Bouikhalene, 2012) tout en tenant compte des caractéristiques suivantes :

- La densité des pixels noirs dans l'image.
- Les niveaux de gris détectés sur l'image.
- Le nombre des lignes noires et blanches (afin de bien différencier les textes des images).

7. Conclusion

La reconnaissance des documents est effectuée par un développement permettant le classement et le tri des contenus. Elle aura comme entrée une image numérisée nettoyée ou une image synthétique et elle est composée de deux étapes successives, une pour la reconnaissance de la structure physique le cas des documents a structure complexe (ou segmentation) et l'autre pour la reconnaissance de la structure logique en utilisant un réseau de neurones.

Références

J. André et V. Quint (2004). Structures et modèles de documents. Technical report INRIA/ IMAG.

A. S. Azokly (1995). *Une approche uniforme pour la reconnaissance de la structure physique de documents composites fondée sur l'analyse des espaces*. Thèse à la faculté des Sciences de l'Université de Fribourg, Suisse.

B. Bouikhalene (2012). *Cours sur Intelligence artificielle*. FST Béni Mellal, Maroc.

B. Caldairou (2012). *Arbre des composantes connexes : méthodologie et application à la segmentation d'images médicales*. Technical report Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection Université Jean Monnet – Saint-Etienne.

N. T. Oanh et S. TABBONE (2004). *Binarisation d'images de documents graphiques*. Technical report Laboratoire Lorrain de Recherche en Informatique et ses Applications à l'Université de Nancy 2.

I. Todinca, J. Tesson (2008). *Cours : Algorithmique des graphes - quelques notes de cours*. Université d'Orléans, France.

TICAM'12, p. 251-260

Hybridation des Modèles de Markov Cachés et de la Logique Floue pour la Reconnaissance des Caractères Tifinaghes Manuscrits

Aissa Haidar Mohamed Fakir Omar Bencharef
Equipe de traitement de l'information et de télécommunications, FST Béni Mellal, USMS.
fakfad@yahoo.fr, bencharef98@gmail.com

Résumé

Dans cet article, nous présentons une approche de reconnaissance des caractères Tifinaghes manuscrits isolés basée sur l'hybridation des modèles de Markov cachés et de la logique floue. L'extraction des caractéristiques est réalisée par le codage de la chaîne de Freeman. Malgré le problème de discontinuité du squelette des caractères manuscrits, l'usage d'une grande base de données (AMHCD) nous a permis de surmonter partiellement ce problème.

1. Introduction

Plusieurs approches ont été utilisées jusqu'à présent pour la reconnaissance des caractères latins (Tlemsani *et al*., 2007), arabes (Fakir *et al*., 2000; Fakir, 2011; Ben Amara, 2000; Soulef, 1993) et chinois (Zheng, 1997; Yang, 2007), mais peu d'articles ont été publiés dans le domaine de la reconnaissance des caractères amazighes (Tifinaghes) .

Le dynamisme de recherche dû aux efforts de l'Institut Royal de la Culture Amazighe (IRCAM) a rendu le sujet de l'informatisation de la culture amazighe un sujet de recherche active qui produit des dizaines d'articles par ans. Actuellement dans la littérature consacrée à la reconnaissance des caractères Tifinaghes, on trouve des approches : stochastiques, statistiques, géométriques et syntaxiques (Amrouch, 2009; Es Saady, 2009; Fakir, 2009; Gounane, 2011; El Ayachi, 2011; El Yachi, 2010; Fakir, 2011; Bencharef, 2011; El Kessab, 2011; Es Saady, 2011).

Dans cet article une approche hybride de reconnaissance est étudiée avec la fusion d'une approche stochastique (Modèle de Markov Caché) et d'une approche statistique (K plus proches voisins floue). La motivation sous-jacente est d'intégrer la logique floue avec l'approche stochastique d'un HMM pour évaluer leur rendement mutuel.

2. Caractéristiques du Tifinaghe

Le Tifinaghe est un alphabet utilisé par les amazighs, essentiellement les touaregs. C'était autrefois un alphabet consonantique. Cet alphabet a subi des modifications et des variations inévitables depuis son origine jusqu'à nos jours. Les caractères Tifinaghes s'écrivent

normalement de gauche à droite et verticalement du haut en bas. Figure 1 illustre les caractères Tifinaghes adoptés par l'IRCAM.

Figure 1 : Caractères Tifinaghes officiels (IRCAM)

3. Modèles de Markov Cachés

3.1. Définition des MMCs

Un modèle de Markov caché est un double processus stochastique (X_t, Y_t) 1<=t<=T. La chaîne interne X_t non observable et la chaîne externe Y_t observable s'allient pour générer le processus stochastique.

La chaîne interne est supposée, pour chaque instant, être dans un état où la fonction correspondante génère une composante de l'observation. La chaîne interne change d'état en suivant une loi de transition. L'observateur ne peut voir que les sorties des fonctions aléatoires associées aux états et ne peut pas observer les états de la chaîne sous-jacente, d'où le terme de Modèles de Markov Cachés (ou Hidden Markov Model) (Belaid, 1997).

Le processus (X_t) 0<=t<=T est une chaîne de Markov d'ordre 1, il doit vérifier :

$$P(X_{t+1} = q_j \mid X_t = q_i, ..., X_0 = q_0) = P(X_{t+1} = q_j \mid X_t = q_i) = a_{ij} \quad (1)$$

Pour tout $t >= 0$

Le processus (Y_t) $0 <= t <= T$, processus observable, vérifie :

$$P(Y_t = y_t \mid X_t = q_i) = b_i(y_t) \quad (2)$$

3.2. Éléments d'un MMC

Selon (Belaid, 1997), un MMC est caractérisé par :

- N, le nombre d'états dans le modèle. On note $S = \{S_1, S_2, ..., S_N\}$, l'ensemble des N états du modèle, q_t l'état au temps t (q_t appartient à S).

- M, le nombre de symboles d'observation par état, c'est-à-dire la taille de l'alphabet discret, les symboles d'observation correspondent aux sorties physiques du système à modéliser. On note $V = \{V_1, V_2, ..., V_M\}$, l'ensemble discret des M symboles, et o_t un symbole au temps t (o_t appartient à V).

- La distribution de probabilité de transition d'un état $A= \{a_{ij}\}$ tel que :

$$a_{ij} = P [q_{t+1} = S_j \setminus q_{tS} = S_i], \ 1 \leq i, j \leq N. \qquad (3)$$

Pour le cas spécial où à partir de tout état, on peut atteindre directement n'importe quel autre état, on a : $a_{ij} > 0$, pour tout i, j. Pour autres types de MMC, on a $a_{ij} = 0$ pour un seul pair (i,j) ou plus.

- La distribution de probabilité de symboles d'observation dans un état j, $B=\{ b_j (k)\}$, où

$$b_j (k) = P[v_k \text{ en } t \setminus q_t = S_j], \ 1 \leq j \leq N, \ 1 \leq k \leq M. \qquad (4)$$

- La distribution d'état initial $\pi = \{\pi_i\}$ tel que : $\pi_i = P [q_1 = S_i] \ \ 1 \leq i \leq N.$ (5)

Donc, la spécification complète d'un MMC nécessite la spécification des deux paramètres du modèle N et M, la spécification des symboles d'observations, et la spécification des trois mesures de probabilités A, B et π. L'ensemble complet des paramètres d'un MMC peut être noté : $\lambda = (A, B, \pi)$.

Dans un MMC, les contraintes (markoviennes) suivantes doivent être respectées :

$$\begin{aligned} &\sum_{i=1}^{N} \pi_i = 1 \\ &\sum_{i=1}^{N} a_{ti} = 1 \\ &\sum_{j=1}^{N} a_{ij} = 1 \quad , \quad 1 \leq i \leq N \\ &\sum_{j=1}^{N} b_j (k) \end{aligned} \qquad (6)$$

3.3. Types des MMCs

Selon sa topologie, un MMC peut être l'un des deux types, ergodique ou gauche-droite.

MMC ergodique

Dans ce type, tout état est directement atteignable depuis tout autre état. Il est plus général et intéressant lorsque le modèle représente un processus dont on veut suivre les évolutions des états.

MMC gauche-droite

Dans ce type, si le temps augmente, alors les indices des états augmentent également. Il est utilisé pour suivre des observations dont l'évolution se fait dans un ordre donné tel que la reconnaissance de la parole.

4. Logique floue

La logique floue est une extension de la logique booléenne proposée par Lotfi Zadeh (Daoudi, 2000). Elle se base sur la théorie des ensembles flous, qui est une généralisation de la théorie des ensembles classiques. Par abus de langage, nous utiliserons indifféremment les termes sous-ensembles flous et ensemble flous. Les ensembles classiques sont également appelés ensembles nets, par opposition à flou, et de même la logique classique est également appelée logique booléenne ou binaire.

4.1. K plus proches voisins floue

Le principe de la méthode KPPV floue est similaire à la règle des KPPV classiques. Il se base sur une étape de recherche des k points voisins les plus proches du prototype à étudier. On examine parmi les K voisins retournés, le taux de mélange des classes. Le KPPV flou permet en plus, d'attribuer suivants les distances aux classes des k prototypes voisins, des degrés d'appartenance à ces classes. Le point à classer x_i se voit attribuer un coefficient d'appartenance $\mu_{j\,i}$ à chaque classe j. Celui-ci est en fonction des distances et des coefficients d'appartenance de ses k plus proches voisins. Ces coefficients doivent vérifier l'appartenance à l'intervalle [0,1] de $\mu_{j\,i}$, pour tous les i et tous les j (Keller, 1985; Bondugula *et al*., 2005)

$$\sum_{i=1}^{c} u_{u} = 1 \quad \text{pour tous les i,} \qquad 0 < \sum_{i=1}^{N} u_{u} \leq N \quad \text{pour tous les j}$$

Les coefficients d'appartenance d'un nouveau point x_i à la classe j est donnée par :

$$u_{ji} = \sum_{I=1}^{k} \frac{u_{jI}\left(\left|x_j - x_I\right|\right)^{\frac{m-1}{2}}}{\sum_{I=1}^{k}\left(\frac{1}{\left(\left|x_j - x_I\right|\right)^{\frac{m-1}{2}}}\right)} \qquad (7)$$

Tel que m est un entier, où $\mu_{j\,i}$ est le coefficient d'appartenance à la classe C_j, $\mu_{j\,i}$ =1 si appartient aux K plus proches voisins $\mu_{j\,i}$ =0 si non, de la $i^{ème}$ observation, parmi les k plus proches voisins de x_i. La variable quant à elle, détermine l'importance de la contribution de la distance dans le calcul de la fonction d'appartenance (contrôle l'efficacité de l'ampleur de la distance). C'est le paramètre de fuzzification : si m croit, la contribution des voisins est d'avantage pondéré et la notion de distance perd de son importance, si m tend vers l'unité, la contribution des voisins les plus proches sera favorisée, ainsi la notion de distance prend de l'importance, si m vaut 2, la contribution de chaque voisin est pondérée par l'inverse de la distance respective, au carré, qui sépare une observation de l'observation à classer.

5. Système de reconnaissance

Notre système de reconnaissance de caractères Tifinaghes manuscrits se base sur une hybridation de deux approches stochastique et statistique. On a en entrée une image d'une lettre manuscrite qui correspond à une lettre de l'alphabet Tifinaghe, cette dernière sera nettoyée par un processus de prétraitement qui comporte trois sous module : un sous module de prétraitement qui se charge de la normalisation afin d'obtenir un mot adapté à une dimension fixée par le système, ensuite nous allons passer cette même image à un autre sous module de binarisation qui se charge de la conversion de l'image en une image bitonale, pour obtenir image nettoyée et enfin nous allons passer cette même image à un autre sous module qui se charge de la squelettisation du mot. Cependant, l'image prétraitée va passer à un sous-système d'apprentissage pour qu'elle puisse être traité, ce dernier se charge d'extraction des caractéristiques importantes par le codage des chaînes de Freeman et après on passe ces caractéristiques (séquences de code de Freeman) aux modèles da Markov cachés via l'algorithme de Baum Walch qui va estimer les paramètres du modèle pour chaque caractère par le critère de maximum de vraisemblance pour un nombre d'itérations fixé par rapport aux paramètres d'un modèle initial défini au départ, on sauvegarde les valeurs des paramètres estimés des nouveaux modèles ainsi les valeurs de vraisemblances, ces valeurs de vraisemblance seront classées à l'aide de KPPV-floue ou KPPV enfin de les enregistrer sur une base de données d'apprentissage.

5.1. Prétraitement

Dans le but d'automatiser la reconnaissance de l'écriture, il faut préparer les images à traiter. Les étapes préliminaires, qui sont nécessaires à la reconnaissance sont la binarisation et la squelettisation.

Binarisation

La binarisation est la première étape de prétraitement, elle consiste à convertir l'image numérisée en une image binaire. Cependant, la binarisation est une opération qui produit deux classes de pixels, en général, elles sont représentées par des pixels noirs et des pixels blancs.

Squelettisation

La squelettisation sert à obtenir une épaisseur égale à 1 du trait d'écriture et de se ramener ainsi à une écriture linéaire. Le squelette doit préserver la forme, connexité, topologie et extrémités du tracé, et ne doit pas introduire d'éléments parasites (Figure 2).

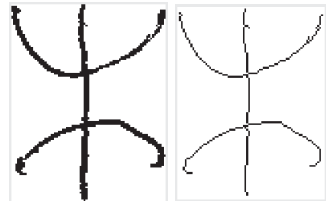


Figure 2 : Exemple de squelettisation d'une image binaire

Nous avons utilisé l'algorithme de squelettisation de Zhang-Suen, l'idée principale de cet algorithme est qu'il élimine tous les points de contour de l'objet sauf les points qui appartiennent au squelette.

5.2. Extraction des caractéristiques

5.2.1. Code de Freeman

La première étape dans cette phase est d'extraire les chaînes de code de Freeman pour chaque image de la base. Le code de Freeman, introduit la première fois en 1961 dans (Freeman, 1961), a pour but de coder le contour d'un objet par :

- une information absolue correspondant aux coordonnées d'un point de départ.
- une chaîne de codants donnant la position relative du point suivant du contour de l'objet selon une des représentations présentées figure 3 (codage utilisant 4 ou 8 directions).

Ainsi, en utilisant 8 directions, le codant 0 signifie que le pixel suivant du contour de l'objet se situe à droite du pixel courant, et le codant 5 désigne un pixel suivant en bas à gauche du pixel courant. Chacune des 4 et 8 directions peut être codée respectivement sur 2 et 3 bits, induisant une forte compression sans perte de l'image.

3	2	1
4	*X*	0
5	6	7

	1	
2	*X*	0
	3	

Figure 3 : Code de Freeman à 4 ou 8 directions.

Pour obtenir un code de Freeman, on sélectionne un point appartenant à la frontière interne de A (par exemple le premier pixel rencontré par balayage électronique), on mémorise les coordonnées de ce premier point (ici, (3,8)), puis on cherche son plus proche voisin appartenant à A (au sens d'un voisinage V4) selon un sens de rotation donné (ici, le sens direct). On réitère ensuite cette dernière opération jusqu'à revenir au point de départ.

Ainsi, de proche en proche, on reconstitue la forme de l'objet A en donnant le codage de Freeman (Freeman, 1974) de la frontière de A.

Cet exemple donne la chaîne 667760700001001012222234444444544545. Un schéma un peu plus détaillé du code obtenu sur cet objet est présenté à la figure 4.

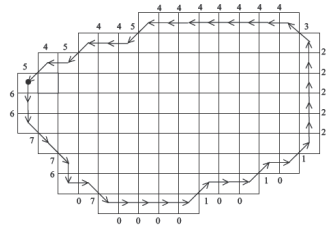


Figure 4 : Exemple d'objet et du code de Freeman à huit directions de son contour.

5.3. Choix du modèle initial

Pour définir l'architecture du modèle, nous devons tenir compte de la topologie et du nombre d'états du modèle (Belaid, 1997). La topologie adoptée dans notre système est de type gauche-droit avec saut inter-états et intra-état. Ce type de modèles a l'avantage de conserver la notion du temps dans la modélisation. En outre, c'est le type le moins gourmand en temps de calcul et en nombre de paramètres à estimer lors de l'apprentissage.

Nous avons considéré les codes de Freeman extraits à partir de l'image de caractères comme des observations pour notre modèle.

5.4. Apprentissage

L'apprentissage des paramètres des modèles HMMs correspondants aux classes de séquences est réalisé par l'algorithme de Baum-Welch, qui va estimer les paramètres du modèle de chaque séquence par la maximisation en plusieurs itérations la valeur de vraisemblance par rapport aux paramètres de modèle initial.

Nous enregistrons les nouveaux paramètres de chaque modèle ainsi la valeur de vraisemblance par rapport au modèle initial, ces valeurs de vraisemblance qui vont nous permettre de classifier les modèles estimés pour les différents caractères.

5.5. Classification

Lors de cette phase, nous allons utiliser la méthode K-plus proche voisins classique et aussi sa version floue pour la classification des valeurs de vraisemblance calculées à l'aide de modèles de Markov cachés à l'étape précédente.

6. Expérimentations

Une série d'expériences ont été réalisées pour évaluer l'efficacité de l'approche proposée. Ces expériences ont été effectuées sur la base de données de caractères amazighes isolés manuscrits (AMHCD) (Es Saady, 2011).

Dans notre expérience, 1600 images de caractères de la base d'AMHCD ont été utilisées pour former les modèles HMMs, et 800 images de caractères ont été utilisées pour tester les performances d'identification. Certains résultats selon le nombre d'états sont énumérés dans le tableau ci-dessous.

	Taille de la base d'apprentissage	400	600	800	1200	1600
	Taille de la base de test	200	300	400	600	800
HMM et KPPV-floue	Taux de reconnaissance	66.12%	73.33%	76.34%	87.78%	88.45%
	Taux d'erreur	33.88%	26.67%	23.66%	12.22%	11.55%
	Temps d'apprentissage (s)	13.899	29.8538	46.9346	55.5231	59.4683
HMM et KPPV classique	Taux de reconnaissance	57.29%	60.18%	69.78%	71.14%	72.55%
	Taux d'erreur	42,71%	39.82%	30.22%	28.86%	27.45%
	Temps d'apprentissage	15.239	29.609	50.899	58.854	65.989

Tableau 1 : Taux de reconnaissance, taux d'erreur et temps d'apprentissage

Ces résultats expérimentaux montrent que l'approche proposée est plus prometteuse que l'approche basée sur les HMM sans la notion de logique floue.

Au niveau des taux de reconnaissance, nous avons obtenu un taux de reconnaissance de 88.45% ce qui présente un excellent taux de reconnaissance pour les caractères manuscrits par rapport aux travaux récemment publiés sur le sujet (Amrouch, 2009; Gounane, 2011; EL Ayachi, 2011; El Yachi, 2010; Fakir, 2011; Bencharef, 2011; El Kessab, 2011 ; Es Saady, 2011).

Pour le problème de discontinuité du squelette des caractères manuscrits, l'usage d'une grande base de données (AMHCD), nous a permis de surmonter partiellement ce problème.

7. Conclusion

Dans cet article, nous avons utilisés les MMC et la logique floue pour la reconnaissance des caractères Tifinagh isolés manuscrits. Nous avons présenté les résultats expérimentaux effectués sur la base de données de caractères amazighs isolés manuscrits (AMHCD). Les résultats obtenus sont prometteuses 88.45% pour MMC et KPPV-floue et 72.55% pour MMC et KPPV en utilisant une base de données de 1600 caractères.

Le travail réalisé nous ouvre plusieurs perspectives. Nous essayerons d'étendre notre approche de reconnaissance des caractères vers la reconnaissance des mots, phrases et de textes.

Références

M. Amrouch, Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass (2009). Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform. *International Conference on Multimedia Computing and Systems* (ICMCS '09).

A. Belaid, G. Saon (1997). Utilisation des processus markoviens en reconnaissance de l'écriture. *Traitement du Signal*, 14(2): 161-177.

N. Ben Amara, A. Belaïd, N. Ellouze (2000). Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : État de l'art. *3ème Colloque international francophone sur l'écrit et le document* (CIFED'00), Lyon, France, pp. 181-191.

O. Bencharef, M. Fakir, B. Minaoui, B. Bouikhalene (2011). Tifinagh Character Recognition Using Geodesic Distances, Decision Trees & Neural Networks. *International Journal of Advanced Computer Science and Applications*, Special Issue on Artificial Intelligence.

R. Bondugula, O. Duzlevski, D. Xu (2005). Profiles and fuzzy k-nearest neighbor algorithm for protein secondary structure prediction. *Asia Pacific Bioinformatics Conference*.

K. Daoudi, D. Fohr, C. Antoine (2000). A new approach for multi-band speech recognition based on probabilistic graphical models. *International Conference on Spoken Language Processing* (ICSLP).

R. El Ayachi, K. Moro, M. Fakir, B. Bouikhalene (2010). On the recognition of tifinaghe scripts. *Journal of Theoretical and Applied Information Technology*.

Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass (2011). AMHCD: A Database for Amazigh Handwritten Character Recognition Research. *International Journal of Computer Applications* 27(4):44-48, New York, USA.

B. El Kessab, C. Daoui, B. Bouikhalene, M. Fakir (2011). Utilisation des réseaux de neurones et le modèle de Markov pour la reconnaissance des caractères Tifinagh manuscrits. *SITACAM,* Agadir, Maroc, 06-07 mai 2011. pp. 31-40.

M. Fakir, C. Sodeyama (1993). Recognition of Arabic printed Scripts by Dynamic Programing Matching Method. *IECICE Trans. Inf&Syst,* vol. E76- D, n° 2.

M. Fakir (2001). Reconnaissance des Caractères Arabes Imprimés. Thèse, pp. 28-36, Faculté des sciences, Université de Semlalia, Maroc.

M. Fakir, O. Bencharef, B. Bouikhalene, B. Minaoui (2011). Tifinagh Character Recognition Using Riemannian Metric, SVM & Neural Networks, *International Journal of Advances in Science and Technology,* 2(6): 1-9.

H. Freeman (1961). On the encoding of arbitrary geometric configurations. *IRE Trans. Electron. Comput.* EC-10, pp. 260-268.

S. Gounane, M. Fakir, B. Bouikhalene (2011). Recognition of Tifinagh Characters Using Self Organizing Map and Fuzzy K-Nearest Neighbor. Thèse de Master, FST Beni Mellal, Maroc.

J. M. Keller, M. R. Gray, J. A. Givens (1985). A Fuzzy K-Nearest Neighbor algorithm. *IEEE transaction on systems, man, and cybernetics*, vol. SMC-15, pp. 580-585.

N. Soulef, F. Nadir (1993). Reconnaissance de l'Ecriture Arabe par Systèmes Flous. Thèse de Master. Département Informatique, Université Badji Mokhtar, pp. 31-37.

R. Tlemsani, A. Benyettou (2007). Application des réseaux bayésiens dynamiques à la reconnaissance en-ligne des caractères isolés. *The 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications,* Tunisie.

D. Yang, L. Jin (2007). Kernel Modified Quadratic Discriminant Function for On-line Handwritten Chinese Characters Recognition, *The 9th International Conference on Document Analysis and Recognition* (ICDAR'07), vol. 1, pp. 38-42.

J. Zheng, X. Ding, Y. Wu (1997). Recognizing Online Handwritten Chinese Character via FARG Matching. *The 4th International Conference on Document Analysis and Recognition* (ICDAR'97), pp. 621-624.

TICAM'12, p. 261-268

Recognition of Amazigh Characters Using Visual Rotation Invariant Features

Younès Raoui **El Houssine Bouyakhf**
LIMIARF, Université Mohamed V-Agdal, Faculté des Sciences, Rabat
raoui@ieee.org
bouyakhf@mtds.com

Abstract: In this paper, we present a new method for Amazigh Character Recognition. It is based on the extraction of salient points from the Tifinagh characters using a visual detector and descriptor. The novelty of this work is, first, the development of a new detector descriptor having a high repeatability which outperforms other transforms like (SIFT, SURF, Harris), second, its application for the recognition of the Amazigh characters. Our method is called Rotation Invariant Detector (RID). We use steerable filters for representing the image in many scales and orientations, then, the Harris detector applied to each of the obtained images, because of its high speed in the detection operation. In order to recognize a character, we match our image test to the database of learnt characters. We present in this work two applications based on RID:

- The visual recognition of handwritten Tifinagh characters using the paradigm (learning-classification).
- The conversion of a handwritten character to a typed character with a computer.

1. Introduction

The computer vision is the science of recognizing objects in the images and interpreting them. During the last decade, this discipline has been widely investigated because of its relation to other fields such as character treatment and image retrieval. In order to interpret an image, the analysis has to move from image processing tasks to object recognition and categorization. Consequently, the algorithms which are used should be faster and more complicated. The description of the images can be obtained with predefined kernels to obtain edges, or by using more recent differential methods to extract coins or feature points. The feature points are better for object recognition. This is because they are more distinctive and repeatable. Harris is one of the first detectors and still gaining success because of its speed (Harris and Stephens, 1998). At the same time, we find also SIFT (Scale Invariant Feature Transform) and SURF (Speeded Up Robust Feature), detectors and descriptors actually very robust (Bay and Tuytelaars, 2006).

To recognize an object in a certain image, first we apply the algorithm which detects and describes the set of the training images. Second, we recognize the object using the matching algorithm in order to compute similarities between the image database and the test image.

We are interested in this work to the optical character recognition. This concept emerged few years ago. It is a popular research area because of its utilization in several important domains such as postal automation and document analysis.

Recently, some efforts have been reported in literature for Amazigh character recognition based on artificial neural networks (Ait Ouguengay and Taalabi, 2009), Hidden Markov Models (Es Saady *et al*., 1998), and Hough transforms. We propose in this paper to use the paradigm of visual detection description with salient points to recognize Tifinagh characters and to convert them to typed characters with a computer. This is important in particular to develop modules which could be installed on popular word processing software using the visual approach.

This paper is organized as following: In the first section, we will introduce the approach of image interpretation with salient features. In the second section, we present some issues about Amazigh language. In the third section, we present a new approach for coin detection and description which we call Rotation Invariant Descriptor (RID). In the fourth section we present application of the developed detector descriptor on handwritten character recognition and conversion to typed characters. Finally, we make some conclusions.

Figure 1: The TIFINAGH Alphabet

2. The Amazigh language

The Amazigh language is spoken in many countries in Africa such as Morocco, Algeria, Tunisia and Egyptian Oasis. In Morocco, the term Berber (Amazigh) includes the three main Moroccan variants: Tarifite, Tamazighte and Tachelhite. More than 40% of the country's population speaks Berber. The establishment of the "Royal Institute of the Amazigh Culture" (IRCAM) carried out a major action to standardize the Amazigh language (Es Saady *et al*., 2011). The IRCAM has as main goal to regulate this language taking as central actions the language and the culture dimensions. Thus, through these procedures, the IRCAM has

succeeded to save this language and to promote it. Till the beginning of the third millennium, big efforts were done in the scholar books in order to officialize this language. The Tifinagh is the writing alphabet of the Amazigh language. It has known many changes from its origin. In North Africa, an old version which dates from the 3rd century BC to the 3rd century AD was spoken in the Berber community.

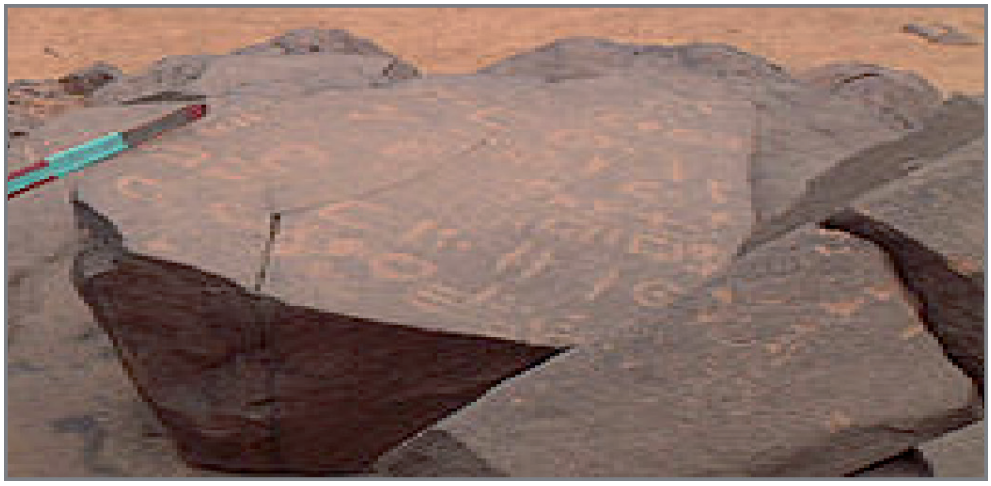


Figure 2: Old Tifinagh script

3. Related works

In the area of visual feature extraction, after the contributions of Moravec and Harris which introduced the concept of corners based on differential operators (Harris and Stephens, 1998), Lowe proposed SIFT (Scale Invariant Features Transforms) to insure some of the properties like invariance to rotations, scales and illumination. This detector descriptor is based essentially on the computation of the differential of Gaussians followed with the computation of the gradient vector around each keypoint. Two years after, (Bay and Tuytelaars, 2006) developed SURF (Speeded Up Robust Feature), another detector descriptor, because SIFT was not enough repeatable, distinctive and robust to do tasks like object recognition and object categorization efficiently.

Nevertheless, these detector descriptors still have to be improved when doing tasks of recognition. Because, this requires having invariant features when the robot changes its viewpoint. In other words, the repeatability should be better than 70% between two successive images of the same scene. We present in the next section, a new method for detector computation based on Harris detector and steerable filters. It gives a repeatability of 90%.

4. Generating local features

In character processing interpretation, one of the main properties on which the computer society is focused is the ability to keep invariance when the angle of view from which we see the character is changing (Mikolajzyk and Schmid, 2004). We propose in this paper a novel method for character detection and description which increases the repeatability rate. Our method is divided into 2 main steps:

- Detection of feature points with an improved variant of the Harris detector.
- Description of the patch around each feature using an estimation of the color frequencies.

4.1. Steerable filters

Steerable filters are orientation selective convolution kernels used for feature extraction (Forsyth and Ponce 2003). The concept of steerable filters presents an efficient scheme to synthesize filters of arbitrary orientation from a linear combination of basis filters which allow steering a filter to any orientation. Consequently the filter output is determined analytically as a function of orientation.

Considering the n^{th} order Gaussian derivative G relative to an arbitrary orientation θ is given the steerable filter response for the n^{th} order Gaussian derivative $G_n(\theta)$ to an arbitrary orientation θ is the steerable filter is given with the following Gaussian distribution G:

$$G1(\theta) = cos(\theta) * G1(0) + sin(\theta) * G1(90)$$

$$G2(\theta) = K_{21}(\theta) * G2(0) + K_{22} * G2(60) + K_{23}(\theta) * G_2(120)$$

$$G_3(\theta) = K_{31}(\theta) * G_3(0) + K_{32}(\theta) * G_3(45) + K_{33}(\theta)$$

$$*G_3(90) + K_{34}(\theta) * G_3(135)$$

$$K_{3i}(\theta) = \frac{1}{4} * 2 * cos(2(\theta - \theta_i)) + 2 * cos(3(\theta - \theta_i))$$

- Detector computation:

We apply the steerable filters described in 3.1 because they permit to represent the image at many levels of scales and rotate the image according to many orientations. Following this step, we apply the Harris detector to have n pyramids (with reference to n orientations). Each of these pyramids has s levels (with reference to s scales). To cluster these features according to their positions, we use the *K-means* method because of its speed and its simplicity. The corresponding feature point is computed with the calculation of the mean of all features in a cluster. Thus, whatever the orientation (or the scale) of the Tifinagh characters, the coins are the same with a very small error as shown in figure 3.

- Descriptor computation:

Around each interest point, we construct a patch, then we sample the pixels in the frequency domain. In other words, we represent the patch in many quantification levels. Our idea to measure the disorder in the patch is to compute the entropy of each of these determined levels. These entropies are invariant to the changing of features such as scale and orientations. Although scale and orientation are not parts of the detector described above.

4.2. Repeatability

This score is used to compute the degree of invariance of the interest points when the viewpoint of the image changes (see figure 4). It consists to detect the interest points found in both images relative to the lowest total number of interest points. However, only the part of the image that is visible in both images is taken into account.

This ratio is defined with the following quotient:

$$\frac{\text{Number of similar features between the two images}}{\text{Number of the detected features in the two images}}$$

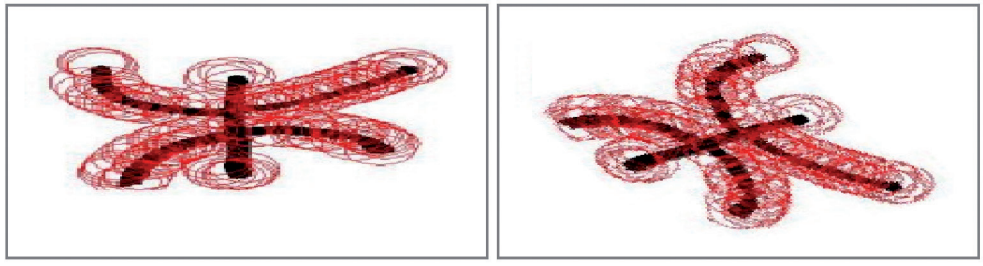


Figure 3: The character written with two persons in different positions and scales

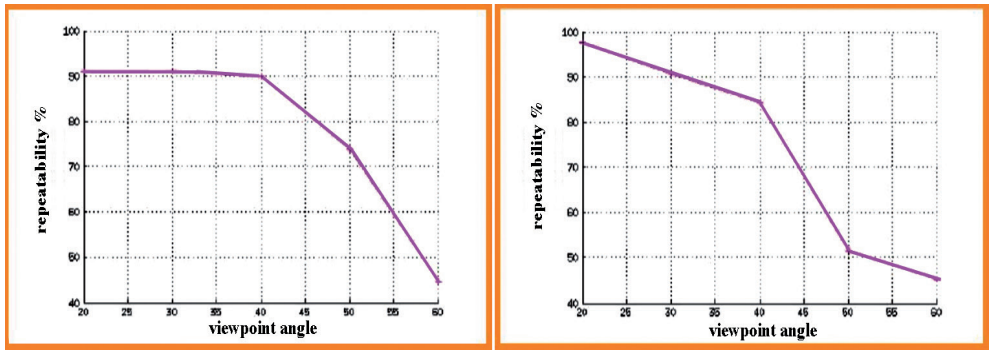


Figure 4: Repeatability computed for 2 characters taken in two different viewpoints

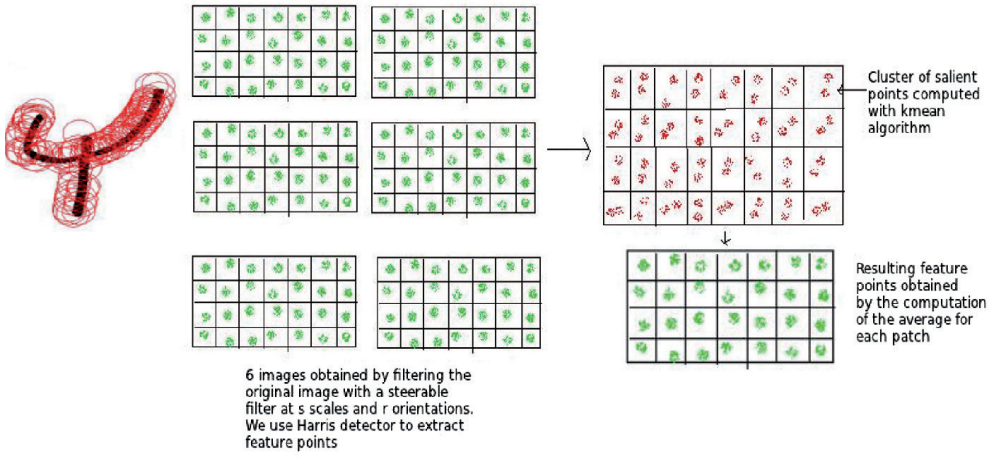


Figure 5: The scheme of the RID detector descriptor

As shown in the first curve of figure 4, the repeatability score varies from an angle of 92% to 74 % for an angle of 50 degree.

In the second curve, the repeatability score varies from 98% for an angle of 20 degree to 55% for an angle of 50.

5. Application of RID to Tifinagh characters recognition

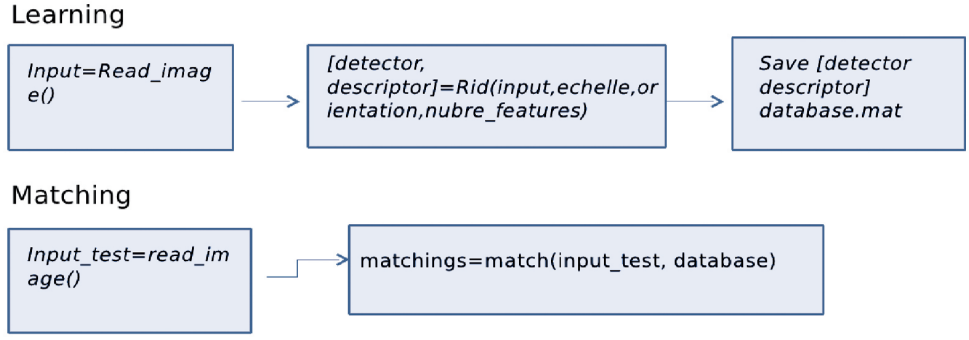


Figure 6: The different modules which we use for the learning of a Tifinagh character. The matching is done with the module match.

As explained before, the two steps for the recognition of a Tifinagh character are: Learning and Matching.

In order to learn a character, we use the database of (Es Saady *et al*., 2011), which contains each of the Tifinagh characters written in several forms. We construct a table of the characteristic of the learnt character for N orientation. Each of the orientations contains all the feature points. Thus, our set contains few hundreds of entries (Es Saady 2012). Upon this features database, we develop two applications:

1. Character recognition: It consists on the matching between this database and the table computed from the test character. To validate the distinctiveness and the efficiency of our method (i.e. RID), we do matching to many databases. Each one contains a different character.

We show in the table 1 that the highest number of correct matches is obtained when associating between the test character and the database of this character (even if they are written differently).

Image database/ Test images			
	66	22,33	41,66
	2,33	49,67	2
	33	10,33	38,33

Table 1: Recognition of handwritten characters: Matching between the image database and the image test using (RID)

2. Conversion to typed characters by computer: This operation is based on the same method as the recognition of handwritten characters. Expect that it doesn't learn the whole forms of the characters. Instead of that, we describe the handwritten character, and we compare it with the visual descriptions of the all typed characters (the alphabet of Tifinagh). The table 2 shows that we can retrieve the correct character from the alphabet. Then with a simple program of conversion, we can display this typed character on the electronic document.

Image database/ Test images	ⵉ	ⵙ	ⵣ	ⵟ
	41,67	6,67	21,33	3
	25,33	49,67	34,67	4
	6	27,67	41,67	3,33
	39,67	36	37,33	35,67

Table 2: Transformation of a handwritten character to the typed character by computer: Matching between the Alphabet and the handwritten character exploits the RID

6. Conclusion

In order to recognize accurately an Amazigh Character, we have proposed in this paper a new detector descriptor invariant to rotations. It allows to retrieve Tifinagh handwritten characters from a database and to retype them in a computer form. Our salient detector descriptor is based essentially on the Harris detector with our stress on the frequencies of colors within the treated character. We show that the repeatability of this new tool out-performs most of the existing features. This work is suitable for the analysis and the treatment of documents written with the Tifinagh language. They could be at the same time used in mobile applications such as Android operating systems to make people aware of this ancient culture.

7. References

Y. Es Saady., A. Rachidi A., M. El Yassa, D. Mammass (2011). AMHCD: A Database for Amazigh Handwritten Character Recognition Research. *International Journal of Computer Applications*, vol. 27, n° 4.

Y. Ait ouguengay, M. Taalabi (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage, *Systèmes intelligents-Théories et applications*.

Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass (2011). Reconnaissance Automatique de l'Ecriture Amazighe à base de Ligne Centrale de l'Écriture. *4ème Atelier international sur l'amazighe et les TIC*, IRCAM, Maroc.

M. Amrouch, A. Rachidi, M. Elyassa, D. Mammass (2010). Handwritten Amazigh character recognition based on hidden Markov models. *ICGST-GVIP Journal*, 10(5):11-18.

C. Harris, M. Stephens (1988). A combined corner and edge detector. *The 4th Alvey Vision Conference*, pp. 147-151.

Y. Es Saady (2012). Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents. Thèse de Doctorat, FSA, Agadir, Maroc.

K. Mikolajzyk, C. Schmid (2004). Scale affine in technology. *IJCV*, vol. 60, pp. 63-86.

D. Forsyth, J. Ponce (2003). Computer Vision: A Modern Approach. Prentice-Hall.

L. G. H. Bay, T. Tuytelaars (2006). SURF: Speeded up Robust Features. *ECCV*. Part I. LNCS, vol. 3951, pp. 404-417. Springer, Heidelberg.

TICAM'12, p. 269-277

Réalisation d'un OCR pour l'Ecriture Amazighe Imprimée

Youssef Es-Saady Mustapha Amrouch Ali Rachidi
Mostafa El Yassa Driss Mammass
Laboratoire IRF-SIC, Université Ibn Zohr B.P. 8106, Hay Dakhla Agadir, Maroc
{essaady2110, amrouch_mustapha}@yahoo.fr,
rachidi.ali@menara.ma,
melyass@gmail.com, mammass@univ-ibnzohr.ac.ma

Résumé

Nous présentons dans cet article une application de reconnaissance automatique de l'écriture amazighe basée sur la ligne centrale horizontale et la ligne centrale verticale du caractère. Après des prétraitements sur l'image d'entrée, le texte est segmenté en lignes et puis en caractères isolés. Les positions des lignes centrales du caractère sont utilisées pour obtenir un ensemble de caractéristiques indépendantes et dépendantes à ces lignes. Ces caractéristiques sont liées aux densités de pixels et sont extraites sur les images binaires des caractères en se basant sur l'utilisation de la technique des fenêtres glissantes. Le système a montré de bonnes performances sur une base de 19437 paternes amazighes et sur la base AMHCD de 20150 caractères amazighes manuscrits.

1. Introduction

La reconnaissance automatique de l'écriture manuscrite ou imprimée reste encore un sujet de recherche et d'expérimentation. Le problème n'est pas encore entièrement résolu bien que l'on sache atteindre des taux assez élevés dans certaines applications et pour certaines langues. Plusieurs recherches scientifiques ont été effectuées sur l'écriture latine, arabe, et autres. Ceci a permis le développement de plusieurs approches de reconnaissance automatique de ces écritures. Par contre, l'écriture amazighe, appelée Tifinaghe, est peu traitée. Quelques travaux ont été menés pour améliorer la situation actuelle (Djematen *et al*., 1998; Ait Ouguengay, 2009; Es-Saady *et al*., 2011; El Yachi *et al*., 2011c; Amrouch *et al*., 2012). Dans ce cadre et dans nos travaux précédents, nous avons proposé un système de reconnaissance de l'écriture amazighe basée sur la ligne centrale horizontal du caractère (Es-Saady *et al*., 2011a). Après l'étape des prétraitements, le texte est segmenté en lignes et en caractères isolés en utilisant les techniques d'analyse d'histogramme de projections horizontales et verticales. Les positions des lignes de base des caractères (une ligne centrale, supérieure et la ligne inférieure du caractère) (*cf*. Figure 1) ont été utilisées pour dériver un ensemble de caractéristiques de densités indépendantes et dépendantes de la ligne centrale horizontale en utilisant la technique des fenêtres glissantes (Elhajj *et al*., 2005; AL-Shatnawi et Khairuddin, 2008; Aida-zade et Hasanov, 2009; Razzak *et al*., 2010). Le système a montré de bonnes performances sur la base de données de 19437 caractères amazighes imprimés développés dans (Ait Ouguengay *et al*., 2009). Les résultats expérimentaux ont montré une amélioration significative du taux

de reconnaissance lors de l'intégration des caractéristiques dépendantes de la ligne centrale horizontale du caractère. Les causes d'erreurs sont principalement dues à la ressemblance entre certains caractères amazighes.

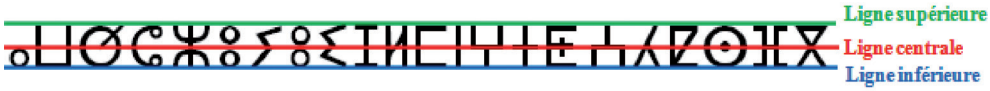


Figure 1 : Les positions des lignes d'écriture sur quelques caractères amazighes

Pour surmonter ces limitations, nous allons ajouter, dans le présent article, d'autres caractéristiques basées sur la ligne centrale verticale du caractère. En effet, la majorité de caractères tifinaghes possèdent la ligne centrale horizontale ou verticale comme un axe de symétrie. Nous citons ci-dessous les caractères amazighes qui possèdent les lignes centrales (horizontale et verticale) comme axes de symétries.

- Les caractères qui possèdent une symétrie orthogonale par rapport à la ligne centrale horizontale du caractère :

  ⵇ, ⵔ, ⵕ, ⵖ, ⵙ, ⵚ, ⵣ, ⵤ, ⵉ, ⵏ, ⵀ, ⵃ, ⵄ, ⵅ, ⵆ, ⵋ ;

- Les caractères qui possèdent une symétrie orthogonale par rapport à la ligne centrale verticale du caractère :

  ⵇ, ⵔ, ⵕ, ⵖ, ⵙ, ⵚ, ⵣ, ⵤ, ⵉ, ⵏ, ⵀ, ⵡ, ⵢ, ⵃ, ⵍ, ⵠ, ⵟ, ⵣ ;

La Figure 2 ci-dessous illustre l'architecture générale du système proposé.

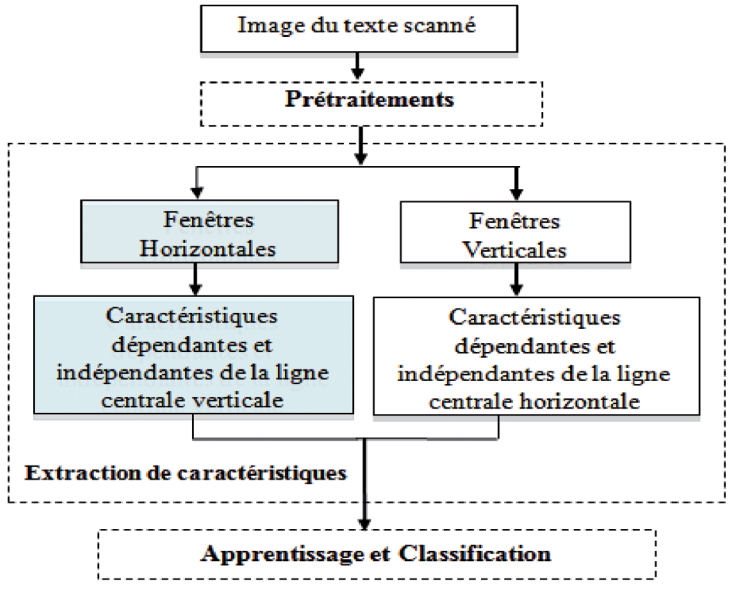


Figure 2 : Architecture générale du système de reconnaissance proposé

Dans la section 2, nous présentons les différentes caractéristiques extraites en utilisant la technique des fenêtres glissantes et les deux lignes centrales (verticale et horizontale). Les résultats expérimentaux seront présentés et commentés dans la section 3. Finalement, une conclusion ainsi que des perspectives futures sont présentées dans la section 4.

2. Extraction des caractéristiques

Dans notre travail précédent (Es-Saady 2011a), les caractéristiques de densités des pixels d'écriture extraites se basent sur la position de la ligne centrale horizontale du caractère. L'image du caractère est balayée de gauche à droite et de haut en bas par une fenêtre glissante qui s'adapte en hauteur à celle du caractère (*cf*. Figure 3). La hauteur et la largeur des fenêtres sont constantes et sont considérées comme des paramètres du système (Elhajj *et al*., 2005).

Dans chaque fenêtre, nous avons généré un ensemble de 19 caractéristiques. Celles ci sont représentatives de densités des pixels d'écriture. L'ensemble de ces caractéristiques extraites comporte 6 caractéristiques qui dépendent de la position de la ligne centrale horizontale, et 13 qui n'en dépendent pas.

2.1. Les caractéristiques indépendantes de la ligne centrale

Pour chaque fenêtre, 13 caractéristiques de densités indépendantes de la ligne centrale sont extraites et sont les suivantes:

- f_1 : la densité des pixels noirs dans la fenêtre ;
- f_2 : le nombre de transitions Noir/Blanc entre cellules ;
- f_3 : la différence de position entre les centres de gravité g des pixels d'écriture dans deux fenêtres consécutives ;
- f_4 à f_{13} : sont les densités de pixels d'écriture dans chaque colonne de la fenêtre.

2.2. Les caractéristiques dépendantes de la ligne centrale

Les caractéristiques dépendantes de la position de la ligne centrale horizontale sont les suivantes :

- f_{14} : la position verticale normalisée du centre de gravité des pixels d'écriture, par rapport à la ligne centrale ;
- f_{15}, f_{16} : les deux primitives qui représentent les densités des pixels d'écriture au dessus et au dessous de la ligne centrale ;
- f_{17}, f_{18} : le nombre de transitions Noir/Blanc entre les cellules situées au dessus et au dessous de la ligne centrale ;
- f_{19} : densité des pixels noirs dans la ligne centrale.

2.3. Extraction des caractéristiques basées sur la ligne centrale verticale

Pour exploiter les informations provenant de la ligne centrale verticale du caractère, nous avons généré un deuxième groupe de caractéristiques de densité dépandantes et indépendantes de cette ligne. Pour créer ce deuxième groupe de caractéristiques, l'image du caractère est

divisée en fenêtres horizontaux (*cf.* Figure 3). Le nombre de fenêtre est constant et il est considéré comme l'un des paramètres du système. Chaque fenêtre est divisée en cellules où la hauteur de cellule est fixe.

Dans chaque fenêtre, nous générons un ensemble de 9 caractéristiques (g_1, g_2, g_3, g_{14}, g_{15}, g_{16}, g_{17}, g_{18}, g_{19}). Ces caractéristiques sont similaires aux caractéristiques basées sur la ligne centrale horizontale (f_1, f_2, f_3, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}), respectivement.

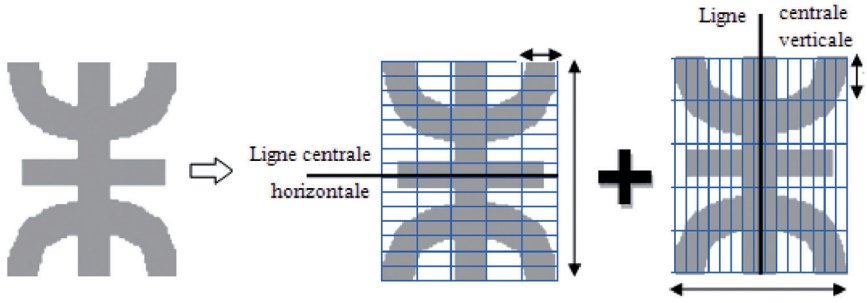


Figure 3 : L'image du caractère est divisée en fenêtres verticales puis en fenêtres horizontales

2.4. L'ensemble de caractéristiques retenues

Le résultat du module d'extraction du vecteur des caractéristiques est maintenant deux groupes de vecteurs de primitives. Chaque groupe correspond au vecteur de caractéristiques extraites à base d'une ligne centrale du caractère. Soient les deux ensembles suivants :

- F : l'ensemble de toutes les caractéristiques, extraites dans chaque fenêtre verticale, basées sur la ligne centrale horizontale

$$F= \{f_1, f_2, \ldots\ldots, f_{19}\}$$

- G : l'ensemble de toutes les caractéristiques extraites dans chaque fenêtre horizontale, basées sur la ligne centrale verticale

$$G= \{g_1, g_2, \ldots\ldots, g_{19}\}$$

Pour optimiser la taille du vecteur de caratéristiques, nous avons éliminé les 10 caractéristiques (f_4, f_5, ……, f_{13}), associées aux densités de pixels d'écriture dans chaque colonne d'une fenêtre verticale et les 10 caractéristiques (g_4, g_5, ……, g_{13}), associées aux densités de pixels d'écriture dans chaque ligne d'une fenêtre horizontale. La raison de diminuer la taille du vecteur de caratéristiques est pour rendre efficace le processus d'apprentissage du réseau de neuronnes. En fait, ces caractéristiques sont indépandantes des lignes de base utilisées et sont moins importantes par rapport aux nouvelles caractéristiques ajoutées. Enfin, nous retenons l'union de deux ensembles F' et G' suivants :

$$F' = F - \{f_4, f_5, \ldots\ldots, f_{13}\} = \{f_1, f_2, f_3, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}\}$$

$$G' = G - \{g_4, g_5, \ldots\ldots, g_{13}\} = \{g_1, g_2, g_3, g_{14}, g_{15}, g_{16}, g_{17}, g_{18}, g_{19}\}$$

L'ensemble de 18 caractéristiques extraites comporte 6 caractéristiques qui dépendent de la position de la ligne centrale horizontale, 6 qui dépendent de la position de la ligne centrale verticale, et 6 qui n'en dépendent pas. Ces caractéristiques alimenteront un réseau de neurones multicouches dans les phases d'apprentissage et de reconnaissance.

3. Résultats et Discussion

Pour évaluer les performances de la méthode proposée et dans un premier temps, nous avons réalisé des expériences sur les deux bases de caractères amazighes : la base des patterns de la graphie amazighe (Ait Ouguengay *et al*., 2009) et la base AMHCD de caractères amazighes manuscrits (Es-Saady *et al*., 2011b). Les tests ont été effectués en fonction de l'intégration des caractéristiques, dépendantes et indépendantes de la ligne de base (Es-Saady *et al*., 2011c; Es-Saady 2012). Dans un second temps, nous avons developpé une intrface graphique MATLAB permettant la reconnaissance des textes amazighes imprimés.

3.1. Evaluation sur les deux bases de données

Nous avons utilisé les techniques de validation croisée 10 fois pour tester et valider la perfermance du système amélioré (Kohavi, 1995). Le Tableau 1 ci-dessous, présente les résultats du système proposé en utilisant la validation croisée 10 fois sur la base des patterns de la graphie amazighe et sur la base de caractères manuscrits.

Les caractéristiques intégrées	Base des patterns de la graphie amazighe		Base de caractères amazighes manuscrits	
	Taille de la base	Taux de Recon	Taille de la base	Taux de Recon
Caractéristiques indépendantes de la ligne centrale	19437 caractères	**88.68 %**	20150 caractères	**84.49 %**
Caractéristiques dépendantes et indépendantes de la ligne centrale horizontale	19437 caractères	**98.49 %**	20150 caractères	**92.23 %**
Caractéristiques dépendantes et indépendantes de la ligne centrale horizontale et verticale	19437 caractères	**99.28 %**	20150 caractères	**96.32 %**

Tableau 1 : Résultats de reconnaissance du système amélioré en fonction des caractéristiques intégrées en utilisant la validation croisée 10 fois

Pour la base des patterns de la graphie amazighe, le taux de reconnaissance est 98,49% lors de l'intégration des caractéristiques basées sur la position de la ligne centrale horizontale et augmente à 99,28% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale vertical.

En comparant le taux de reconnaissance obtenu par ce système amélioré avec celui obtenu par le système de base, présenté dans (Es-saady *et al*., 2011a), nous constatons une amélioration due à l'intégration des caractéristiques basées sur la position de la ligne centrale vertical. Cela démontre que les caractéristiques basées sur la position des lignes centraux (verticale et horizontale) offrent une amélioration significative à la performance de reconnaissance.

En analysant la matrice de confusion obtenue par le système amélioré sur la base des patterns de la graphie amazighe, nous constatons que seulement quatre lettres de l'ensemble de l'alphabet avaient peu d'erreurs.

La première erreur est due à la ressemblance entre certaines lettres dans différentes polices (le problème de resemblance entre les deux caractères yan (ⵍ) et yaj (ⵊ). Dailleurs, le format du caractère yan (ⵍ) sur la fonte 'tassafut' ressemble entièrement au caractère yaj (ⵊ). Ces erreurs ne peuvent pas réellement être corrigées car même un humain ne pourrait différencier certains de ces lettres. La deuxième erreur consiste à 24 remplacements de la lettre ya (ⵄ) par la lettre yar (ⵔ) sur 627 examples. Ce problème est dû au création des caractères ya (ⵄ) dans la base de données utilisée. La troisième erreur consiste à 13 remplacements de la lettre yazz (ⵣ) par la lettre yaz (ⵣ). Ce problème dû à la grande similarité morphologique entre ces deux lettres. La seule différence entre ces deux lettres est un trait dans le centre de la lettre. La lettre yaz (ⵣ) est une dérivation de la lettre yazz (ⵣ).

Dans le Tableau 3 ci-dessous nous dressons la matrice de confusion obtenue sur la base de caractères amazighes manuscrits. L'étude de cette matrice de confusion a montré que la plupart des erreurs sont principalement dues à la ressemblance entre certains caractères amazighes. Par exemples, des confusions entre les deux lettres yaz (ⵣ) et yazz (ⵣ), entre les deux lettres yatt (ⵟ) et yadd (ⴹ), entre les deux lettres yaz (ⵣ) et yax (ⵠ) et entre les deux lettres yey (ⵢ) et yu (ⵓ).

ⴰ	ⴱ	ⴳ	ⴷ	ⴹ	ⴻ	ⴼ	ⴳ	ⴵ	ⴷ	ⵀ	ⵃ	ⵄ	ⵅ	ⵇ	ⵉ	ⵊ	ⵍ	ⵎ	ⵏ	ⵓ	ⵔ	ⵕ	ⵖ	ⵗ	ⵙ	ⵚ	ⵛ	ⵜ	ⵟ	ⵡ	ⵢ	ⵣ	ⵥ	ⵯ	
ya	**yab**	**yach**	**yad**	**yadd**	**yae**	**yaf**	**yag**	**yagh**	**yah**	**yahh**	**yaj**	**yak**	**yal**	**yam**	**yan**	**yax**	**yar**	**yarr**	**yas**	**yass**	**yat**	**yatt**	**yaw**	**yaq**	**yay**	**yaz**	**yazz**	**yey**	**yi**	**yu**	**Taux d'err %**	
649	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**0,15**	ⴰ
0	626	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	7	3	0	6	0	1	1	0	0	1	0	1	**3,69**	ⴱ
0	2	625	0	0	0	0	1	0	0	0	1	1	0	2	0	0	4	7	1	2	0	1	0	0	0	0	1	0	0	2	**3,85**	ⴳ
0	0	0	643	0	1	0	0	0	0	2	0	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**1,08**	ⴷ
0	0	1	0	599	0	0	0	0	0	0	1	3	0	9	0	0	0	0	0	0	0	29	0	2	0	0	0	0	6	0	**7,85**	ⴹ
1	0	0	4	0	638	0	0	0	0	3	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	**1,85**	ⴻ
0	0	0	0	0	0	635	1	1	0	0	1	0	1	0	0	3	0	0	3	0	0	0	1	0	0	1	2	1	0	0	**2,31**	ⴼ
0	0	0	1	0	0	1	641	0	0	0	0	3	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	1	**1,38**	ⴳ
0	0	0	0	1	0	0	0	639	0	0	0	0	3	0	0	0	0	0	0	4	1	0	1	0	0	1	0	0	0	0	**1,69**	ⴵ
0	0	0	0	0	0	0	1	0	627	0	2	1	3	1	0	0	1	6	3	1	1	0	0	3	0	0	0	0	0	0	**3,54**	ⴷ
0	0	0	2	0	2	0	0	0	0	644	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0,92**	ⵀ
0	0	0	0	0	0	0	4	0	1	0	628	0	0	6	3	1	0	0	1	0	2	0	0	0	0	3	0	0	0	1	**3,38**	ⵃ
0	0	0	1	1	0	0	6	0	3	6	0	623	0	0	0	0	0	0	1	2	0	1	0	6	0	0	0	0	0	0	**4,15**	ⵄ
0	0	0	0	0	0	0	0	1	2	1	0	0	640	0	0	1	0	0	1	3	0	0	0	0	0	1	0	0	0	0	**1,54**	ⵅ
0	0	4	0	3	0	0	0	0	0	0	1	0	0	640	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	**1,54**	ⵇ
0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	644	0	0	0	0	0	2	1	1	0	0	0	0	0	0	0	**0,92**	ⵉ
0	0	0	0	0	0	1	2	2	0	5	0	0	1	0	0	622	0	0	0	0	2	0	0	0	0	13	2	0	0	0	**4,31**	ⵊ
0	2	5	0	2	0	1	0	0	1	0	0	0	0	2	1	0	620	2	2	9	0	0	2	0	0	0	0	0	0	1	**4,62**	ⵍ
0	1	4	0	0	2	2	0	1	4	0	0	3	0	0	0	0	5	617	6	1	1	0	0	0	2	0	1	0	0	0	**5,08**	ⵎ
0	1	2	0	2	0	0	0	1	2	0	0	1	0	0	0	0	6	7	611	6	0	4	0	4	0	0	2	0	0	1	**6,00**	ⵏ
0	2	1	0	0	0	2	0	0	3	2	1	0	1	0	0	2	8	0	9	614	0	1	1	3	0	0	0	0	0	0	**5,54**	ⵓ
0	0	0	0	3	2	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	634	1	0	0	0	0	3	4	0	0	**2,46**	ⵔ
1	9	0	0	23	0	1	0	0	0	0	2	3	0	1	2	0	0	1	1	1	0	589	1	7	0	0	1	0	4	3	**9,38**	ⵕ
0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	648	0	0	0	0	0	0	0	**0,31**	ⵖ
0	0	0	1	1	0	0	0	0	3	0	2	4	0	1	1	0	1	0	2	1	0	0	0	633	0	0	0	0	0	0	**2,62**	ⵗ
0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	644	0	0	0	1	0	**0,92**	ⵙ
0	0	0	0	0	3	1	1	0	0	1	4	0	3	0	0	18	0	0	3	3	2	0	0	0	0	577	33	1	0	0	**11,23**	ⵚ
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	6	0	0	0	2	2	0	0	1	1	40	593	1	0	1	**8,77**	ⵛ
0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	3	0	0	1	0	0	2	0	0	0	3	1	2	621	4	10	**4,46**	ⵜ
0	0	1	0	3	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	1	0	5	635	1	**2,31**	ⵟ
0	0	1	0	0	0	0	3	0	1	0	2	0	0	0	1	1	0	0	0	0	0	2	0	1	1	0	2	9	0	626	**3,69**	ⵡ

Tableau 3 : Matrice de confusion obtenue par le système amélioré, évalué sur la base de caractères amazighes manuscrits

3.2. Interface graphique de l'application

La figure 4 ci-dessous montre un aperçu de l'application développée. En effet, le bouton 'Charger l'image' permet de charger l'image du texte à reconnaitre. Ainsi, le bonton 'Reconnaitre' permet de reconnaitre le texte en appliquant une serie des prétraitements afin d'isoler le caractère et en utilisant la méthode d'extraction de caractéristiques présentée ci-dessus. Pour la classification, nous avons appliqué la corrélation entre la lettre extraite et les lettres modèles préparées en utilisant la base des patterns de la graphie amazighe (Ait Ouguengay *et al*., 2009). Les résultats obtenus sur quelques images des textes amazighes imprimés, montrent la pertinence de l'approche adoptée.

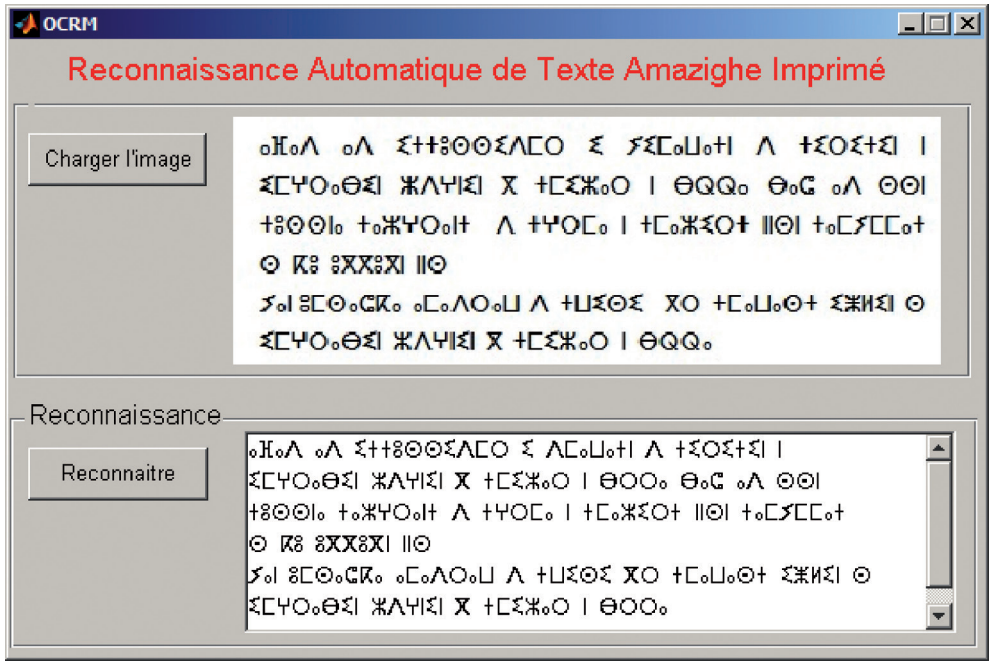


Figure 4 : Interface de l'application testée sur un texte amazighe

4. Conclusion et perspectives

Dans ce chapitre, nous avons présenté un système pour la reconnaissance automatique de l'écriture amazighe à base de la position des lignes centrales de chaque caractère. Plusieurs caractéristiques ont été étudiées et comparées. L'importance de l'utilisation de la position de la ligne centrale horizontale et la ligne centrale verticale du caractère a été prouvée. Les caractéristiques extraites sont basées sur la densité des pixels dérivée dans une fenêtre glissante. Le système développé a été expérimenté sur deux base : une base des patterns de la graphie amazighe et une autre base de caractères amazighes manuscrits développée localement. L'analyse des résultats obtenus par le système montre une amélioration significative du taux de reconnaissance lorsqu'on intègre les caractéristiques dépendantes de la ligne centrale horizontale. La valeur du taux de reconnaissance croit encore lorsqu'on intègre les caractéristiques basées sur la ligne centrale verticale du caractère. Ce qui montre que l'amélioration proposée apporte de bons résultats. Parmi les travaux futurs de ce travail, nous allons ajouter d'autres caractéristiques qui améliorent les résultats pour certains caractères dont le taux de reconnaissance est faible par apport aux restes, telles que les informations sur l'inclinaison possible de l'écriture manuscrite. En plus, nous allons appliquer notre approche sur des documents amazighes.

Références

R. Aida-zade, Z. Hasanov (2009). Word base line detection in handwritten text recognition systems, *International Journal of Electrical and Computer Engineering*, 4(5):310-314.

A. AL-Shatnawi, O. Khairuddin (2008). Methods of Arabic Language Baseline Detection – The State of Art, *IJCSNS International Journal*, 8(10): 137-143.

A. Djematen, B. Taconet, A. Zahour (1997). A Geometrical Method for Printing and Handwritten Berber Character Recognition. Actes de ICDAR'97, pp. 564.

R. El-Hajj, L. Likforman-Sulem, C. Mokbel (2005). Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, *ICDAR'05*, Seoul, Corée du Sud.

R. Kohavi (1995). A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*.

M. Amrouch, Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass (2012). A New Approach Based On Strokes for Printed Tifinagh Characters Recognition Using the Discriminating Path-HMM. *International Review on Computers and Software*, vol. 7, n° 2, ISSN 1828-6003, March 2012.

R. EL Ayachi, M. Fakir, B. Bouikhalene (2011). Recognition of Tifinaghe Characters Using a Multilayer Neural Network. *International Journal Of Image Processing*, vol. 5, Issue 2, 2011.

M. Razzak, M. Sher, S. Hussain (2010). Locally baseline detection for online Arabic script based languages character recognition. *International Journal of the Physical Sciences*, 5(7): 955-959.

Y. Ait Ouguengay, M. Taalabi (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage. Systèmes intelligents-Théories et applications, Paris : Europia, ISBN-102909285553.

Y. Es-Saady (2012). Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents amazighs. Thèse de doctorat, Université Ibno zohr- Agadir, Maroc.

Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass, (2011). Reconnaissance Automatique de l'Ecriture Amazighe à base de Ligne Centrale de l'Écriture. *4ème Atelier international sur l'amazighe et les TIC*, Rabat, Maroc.

Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass (2011). AMHCD: A Database for Amazigh Handwritten Character Recognition Research. *International Journal of Computer Applications*, 27(4): 44-49, ISBN: 978-93-80864-53-2, August 2011.

Y. Es-Saady, A. Rachidi, M. El Yassa, D. Mammass (2011). Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character". *SERSC International Journal of Advanced Science and Technology*, vol. 33, August, 2011, ISSN 2005-4238, pp.33-50.

Study of Effect of Regularized Neural Network on the Accuracy of Handwriting Recognition

Meena M. Makary **Inas A. Yassine**

Systems and Biomedical Engineering Department,
Faculty of Engineering, Cairo University, Giza, Egypt
{meena.magharious, iyassine}@eng.cu.edu.eg

Abstract

The distinction between different handwritten digit recognition become a very important task in our daily usage in tablets, scanner. With the growing availability of powerful computers and diversity of learning algorithms, hand written recognition is determined almost exclusively by the optimization techniques used to increase classification accuracy. In this paper, we proposed the usage of artificial Neural Networks (ANN) based algorithm after adding a regularization term to the cost function. The proposed ANN can correctly identify the handwritten digit and successes to solve the overfitting problem. The reported results are competitive to the highly performed reported results.

1. Introduction

The simultaneous availability of inexpensive powerful computers, powerful learning algorithms and large databases, has caused rapid progress in handwriting recognition in the last few years. Handwritten digit recognition is really a sub problem of handwritten character recognition where an algorithm is needed not only to classify digits, but letters as well. The problem of handwritten digit recognition has been an open problem in the field of pattern classification and of great importance in industry. The heart of the problem lies within the ability to design an efficient algorithm that can recognize digits written and submitted by users via a tablet, scanner, and other digital devices. While recognizing individual digits is only one of many problems involved in designing a practical recognition system, it is an excellent benchmark for comparing shape recognition methods.

Handwriting recognition algorithms are mainly based on one of two methodologies: Memory based and Learning based algorithms. Memory based algorithms stores the training set of images or their patterns with their corresponding labels and classify a new unknown digit by comparing it to each of the stored patterns as K-Nearest Neighbor (K-NN) algorithms. Algorithms based on Euclidean distances, such as the K-NN algorithm, presumably suffering from the «curse of dimensionality» and require a number of training examples that grows exponentially with the dimensionality of the input. Instead of storing the training set, learning based algorithms try to learn from the known patterns and build a classification function

accordingly. An example of a learning based algorithm is a neural network which has proven to be very successful in the handwritten digit recognition (Liu, 2003).

Artificial Neural Networks (ANN) is considered to be one of the most widely used techniques for image recognition (Kussul, 1994; Ososkov, 2003; Lee, 1991). When considering the ANN architecture, recent research recommends the careful design while dealing with multilayer neural networks with local ”receptive fields” and shared weights that may be unique in providing low error rates on handwritten digit recognition tasks (Lee, 1991).

The goal of our study is to design an ANN based algorithm that can correctly identify the digits and free of overfitting problem. Overfitting mainly occurs in case of perfect fitting of training data set and failing to generalize in the unseen test data. However, we are introducing adding a new regularization term which solves this problem by keeping all features but reducing magnitude/values of ANN parameters. Which leads to a simpler hypothesis calculation and a less prone cost function becomes to overfitting. Experimental study results shows promising increase in the accuracy of neural networks.

2. Dataset and Literature Review

2.2. Literature Review

Several authors proposed different algorithm in order to solve the separate digit handwritten problem recognition for MNIST database. Some were based on extracting gradient features on this database (Dong, 2001), whereas others extracted features from gray-scale images (Teow, 2002). Except (Dong, 2001; Belongie, 2002), all the results were obtained on normalized images without heuristic feature extraction. Multilayer ANN was also introduced, in which the local receptive fields function as trainable feature detectors. The features, proposed by Teow et *al*. (Teow, 2002), gave a best accuracy 99.41% using trio-wise linear support vector classifier on direction and stroke end features. Also, Cheng-Lin Liu et *al*. gave the best accuracy 99.58% using SVCs (Liu, 2003). The accuracy of these previously discussed algorithms is listed in Table 1.

Method	Accuracy (%)
LeNet-4 (Y. LeCun, 1995)	98.90
LeNet-5 (Y. LeCun, 1995)	99.05
SVC-poly (C.J.C. Burges, 1997)	98.60
Virtual CV (C.J.C. Burges, 1997)	99.00
Pairwise SVC (Kreβel, 1999)	98.48
Dong et al. (J.X. Dong, 2001)	99.01

Mayraz et al. (G. Mayraz, 2002)	98.30
Belongie et al. (S. Belongie, 2002)	99.37
Teow et al. (L.-N. Teow, 2002)	99.41
Cheng-Lin Liu et al. (Liu, 2003)	99.58

Table 1: Algorithms and their recognition accuracies

3. Materials and Methods

3.1. Data set

The dataset used in our study is The MNIST (LeCun, 1995) which is the most widely used benchmark for isolated handwritten digit recognition. The digits have been size-normalized and centred in a fixed-size image. The database is formed of grayscale images sized 20x20 pixels. Samples of the dataset images are shown in Figure 1.

Figure 1: Sample Images of MNIST data

In our experimental study, 5000 training samples were used, as training set. Each pixel is represented by a floating point number indicating the grayscale intensity at that location. The 20 x 20 grid of pixels is unrolled into a 400-dimensional vector. Each of these training examples becomes a single row in our data matrix X, which forms by the end 5000 x 400 matrix.

$$X = \begin{bmatrix} -\left(x^{(1)}\right)^T- \\ -\left(x^{(2)}\right)^T- \\ \vdots \\ \vdots \\ \vdots \\ -\left(x^{(m)}\right)^T- \end{bmatrix}$$

The labels of each of the training set are then put in vector y sized 5000-dimensional vector y that contains labels for the training set.

3.2. Algorithm Description

The proposed algorithm can be summarized starting with randomly initializing the ANN weights, followed by implementing forward propagation to get $h_\theta(x)$ for any $x^{(i)}$ and computing the cost function $J(\theta)$ defined by equation (1). The back-propagation loop is then used to compute partial derivatives which are minimized using gradient descent or any other advanced optimization method. The gradient checking to compare $\frac{\partial}{\partial \Theta^{l}_{j,k}} \mathrm{j}(\Theta)$ computed using back-propagation versus using numerical estimate of gradient of $j(\Theta)$. The algorithm flowchart is shown in Figure 2.

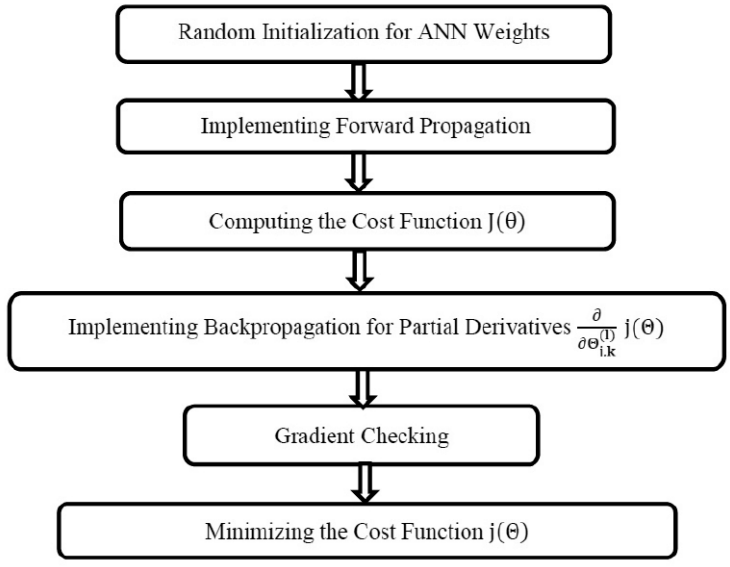


Figure 2: Algorithm Flowchart

Due to the importance of random initialization of ANN parameters for symmetry breaking, we propose the usage of an effective strategy based on the random selection of values for $\Theta^{(1)}$ uniformly in the range $[-\in_{init}, \in_{init}]$.

The choice of $\in_{init}$ is based on the number of units in the network. Which can be calculated using the formula $\in_{init} = \frac{\sqrt{6}}{\sqrt{L_{in} + L_{out}}}$, where L_{in} and L_{out} are the number of units in the layers adjacent to $\Theta^{(1)}$.

3.2.1. Neural Network and Feedforward

Neural network Algorithms is designed to mimic the brain. It was very widely used in 80s and early 90s and now it becomes a state of art technique for many applications as image classification as one of the most common fields. The input to the network is typically a fix-sized, gray-scale, pixel image of the character and no other feature information is necessary (Lee, 1991). However, a neural network with only one hidden layer is the most common and achieved a great performance and low processing time (LeCun, 1995), the inputs to the network are the pixel values of digit gray scale images. Since the images are of size 20 x 20, the No of inputs to the neural network is 400 inputs and 10 output nodes representing the 10 digits.

The cost function used can be defined as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)}\log\left(\left(h_\theta\left(x^{(i)}\right)\right)_k\right) - (1-y_k^{(i)})\log\left(1-\left(h_\theta\left(x^{(i)}\right)\right)_k\right)\right] \qquad (1)$$

where the total number of possible labels $K=10$ and m is the number of training examples.

To compute each element in the summation, $h_\theta(x^{(i)})$ is computed for every example i, where $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ and $g(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

3.4. Regularization

Regularization is an essential technique to improve generalization of neural networks. Traditionally, regularization is conducted by including an additional term in the cost function of a learning algorithm. Having too many features, the learned hypothesis may fit the training set very well, the cost function $J(\theta) = 0$, but fail to generalize to new testing data. To address this overfitting problem, one can reduce the number of features by manually selecting features to keep, but sometimes each of these features contributes a bit to predict the output and this method affects the classification accuracy. Regularization solves this problem by keeping all features but reduce magnitude/values of parameter and by doing so, we will get a simpler hypothesis and the cost function becomes less prone to overfitting.

The regularized cost function can then be defined as (2) :

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)}\log\left(\left(h_\theta\left(x^{(i)}\right)\right)_k\right) - (1-y_k^{(i)})\log\left(1-\left(h_\theta\left(x^{(i)}\right)\right)_k\right)\right] + \frac{\lambda}{2m}\left[\sum_{j=1}^{25}\sum_{k=1}^{400}\left(\left(\Theta_{j,k}^{(1)}\right)^2\right) + \sum_{j=1}^{10}\sum_{k=1}^{25}\left(\left(\Theta_{j,k}^{(2)}\right)^2\right)\right]$$

where λ is the regularization parameter.

3.5. Backpropagation Network

The back-propagation neural network has a feed-forward structure in which nodes in the hidden layer have local "receptive fields" which receive inputs from a limited number of nodes in the layer below. Within a hidden layer, nodes are grouped to form various "feature maps." Nodes of the same feature map share the same set of weights but cover different spatial locations. Each node in the output layer represents one class. Classification is determined by the node with the highest activation function. The learning algorithm of the back-propagation algorithm that minimizes the regularized cost function of the parameters θ.

The intuition behind the back-propagation algorithm can be listed as follows. Given a training example $(x^{(t)}, y^{(t)})$, a "forward pass" is first run in order to compute all the activations throughout the network, including the output value of the hypothesis $h_\theta(x)$. For each node j in layer l, an "error term" $\delta_j^{(l)}$ is computed in order to measure how much that node was responsible for any errors in our output.

For an output node, the difference between the network's activation and the true target value can be directly measured, defined $\delta_j^{(3)}$ (since layer 3 is the output layer). For the hidden units, $\delta_j^{(l)}$ is computed based on a weighted average of the error terms of the nodes in layer $l+1$.

3.6. Numerical Gradient Checking

The cost function $J(\Theta)$ is minimized to perform gradient checking on ANN parameters, through the usage of the gradient checking procedure. Suppose having a function $f_i(\theta)$ that purportedly computes $\frac{\partial}{\partial\theta_i}\, j(\theta)$. Checking if $f_i(\theta)$ outputting the correct derivative values, is needed.

Let $\theta^{(i+)} = \theta + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon \\ \vdots \\ 0 \end{bmatrix}$ and $\theta^{(i-)} = \theta - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon \\ \vdots \\ 0 \end{bmatrix}$

So, $\theta^{(i+)}$ is the same as θ, except its i^{th} element has been incremented by ε . Similarly, $\theta^{(i-)}$ is the corresponding vector with the i^{th} element decreased by ε.

It can be numerically verified that $f_i(\theta)$'s correctness by checking, for each i, that:

$$f_i(\theta) \approx \frac{j(\theta^{(i+)}) - j(\theta^{(i-)})}{2\varepsilon}$$

By assuming $\varepsilon = 10^{-4}$, we will usually find that the left- and right-hand sides of the above will agree to at least 4 significant digits (and often many more).

4. Results and Discussion

The effect of the number of Iteration on the accuracy was firstly studied, while keeping the regularization parameter λ constant. From the results, it's obvious that by increasing the number of iterations, the accuracy increases accordingly as shown in Table 2. It has to be noted here that the processing time increases by increasing the number of iterations. The effect of the regularization parameter was studied by using a constant number of iterations and changing the value of λ. From the results, it can observed that the highest accuracy occurs at $\lambda = 0{,}3$.

No. of Iterations	Accuracy (%) (With lambda (Regularization Parameter is constant = 1))
50	**95.06**
100	97.82
200	**99.20**
400	99.44

Table 2: Effect of the number of iterations on accuracy

Table 3 is listing the results studying the effect of regularization parameter on accuracy. In order to study the effect of number of iterations on the accuracy, the value of the regularization parameter giving the highest accuracy was used for different number of iterations. In order to study the error stability of the regularized ANN algorithm, the cost function for each iteration is calculated for both the proposed regularized ANN algorithm and conventional ANN. Figure 3 shows that both regularized and conventional ANN algorithm cost function are pretty stable through all iterations.

Lambda	Accuracy (With no. of iterations constant = 50)
1	**95.06**
0.9	95.98
0.5	**96.22**
0.2	96.58
0.3	**96.8**
0.1	95.00
0.01	**95.10**

Table 3: Effect of regularization parameter on accuracy

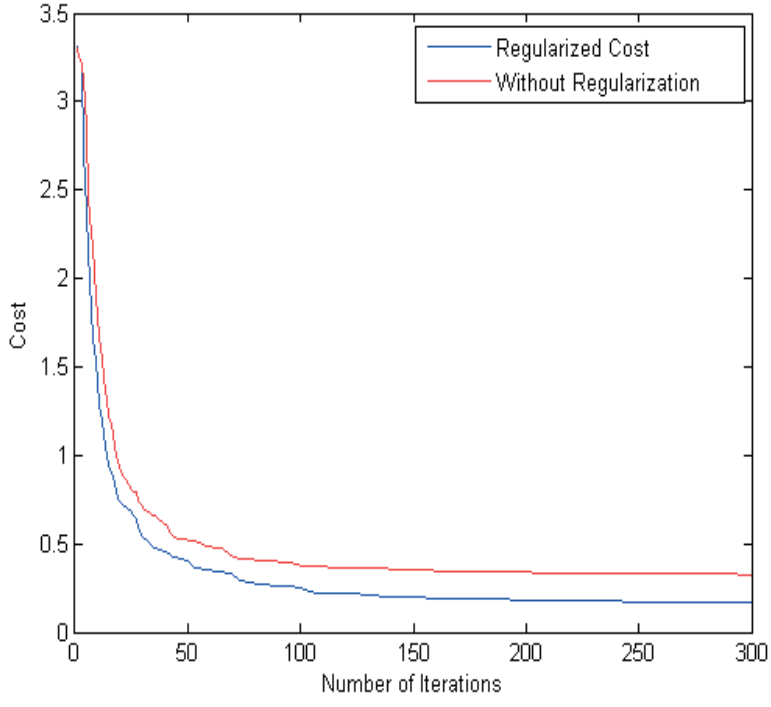


Figure 3: The Cost calculation at each iteration for both conventional and regularized ANN

As shown in Table 4, the accuracy reached 99.98% in case of 300 iterations which can be classified as a very promising result compared to ANN results proposed in the literature. It can be even classified as one of the highest accuracy methodologies compared to results by most of classification techniques, which makes the proposed results of the regularized ANN algorithm very competitive.

Lambda	No. of Iterations	Accuracy (%)
0.3	**200**	**99.84**
0.3	250	99.90
0.3	**300**	**99.98**

Table 4: Effect of the number of iterations on accuracy for λ = 0,3

5. Conclusion

In this study, we presented adding a regularization term for the ANN cost function in order to solve the ovefitting problem. The results, reported in this paper, show that ANN learning accuracy, after adding the regularization term, can achieve accuracy of 99.98%. The accuracy of the proposed algorithm is by far the best when compared to the previously reported results of neural networks as well as that reported for most of the best of handwriting digits recognition classification techniques.

References

Burges C.J.C., Scholkopf B. (1997). Improving the accuracy and speed of support vector learning machines. *Advances in Neural Information Processing Systems*, pp. 375-381.

Kussul E.M., Baidyk T. (1994). Neural random threshold classifier in OCR application. *The Second All-Ukrainian International Conference*, pp. 154-157.

Mayraz G., Hinton G. (2002). Recognizing handwritten digits using hierarchical products of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 2, pp. 189-197.

Dong J.X., Krzyzak A., Suen CY. (2001). A multi-net learning framework for pattern recognition. *The Sixth International Conference on Document Analysis and Recognition*, pp. 328-332.

Kressel U.H.-G. (1999). Pairwise classification and support vector machines. *Advances in Kernel Methods: Support Vector Learning*, pp. 255-268.

Teow L.-N., Leo K.-F. (2002). Robust vision-based features and classification schemes for off-line handwritten digit recognition. *Pattern Recognition*, vol. 11, pp. 2355-2364.

Lee Y., (1991). Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function, and Backpropagation Neural Networks. *Neural Computation*, vol. 3, pp. 440-449.

Liu, C.-L., Nakashima K., Sako H., Fujisawa H. (2003). Handwritten digit recognition: benchmarking of state-of-the-art techniques. *Pattern Recognition*, 36(10): 2271-2285.

Ososkov G., (2003). Effective neural network approach to image recognition and control. *International Conference on Physics and Control*, vol. 1, pp. 242-246.

Belongie S., Malik J., Puzicha J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 4, pp. 509-522.

Le Cun Y., Jackel L.D., Bottou L., Brunot A., Cortes C., Denker J., Drucker H., Guyon I., Muller U. A., Sackinger E., Simard P., Vapnik V. (1995). Comparison of learning algorithms for handwritten digit recognition. *International Conference on Artificial Neural Networks*, pp. 53-60.

Nouvelle Méthode de Reconnaissance Automatique de Caractère Tifinaghe

Mustapha Amrouch Youssef Es-Saady Ali Rachidi
Mohamed El Hajji Mostafa El Yassa Driss Mammass
IRF-SIC Laboratory,
University Ibn Zohr, Agadir, Maroc
{amrouch_mustapha, essaady2110, hajjimohmed}@yahoo.fr, rachidi.ali@menara.ma,
melyass@gmail.com, mammass@univ-ibnzohr.ac.ma

Résumé

Ce papier présente une nouvelle méthode pour la reconnaissance automatique hors ligne de caractères Tifinaghes imprimés. La méthode proposée est basée sur un chemin discriminant (DP-HMM) opérant sur un vocabulaire de base formé de différents graphèmes fondamentaux. Ce vocabulaire est généré en exploitant la redondance de certains traits dans les tracés du caractère Tifinaghe. Un seul modèle HMM globale construit et entraîné sur les éléments du vocabulaire proposé par des primitives structurelles et géométriques. Chaque chemin au long de ce treillis représente une séquence de segments, de ce fait un caractère de l'alphabet Tifinaghe. Pour se faire, les caractères d'entrées sont pré-classés en plusieurs classes morphologiques, chaque classe subira un traitement approprié afin de faciliter la localisation de leurs points d'intérêts et leurs segments. La reconnaissance s'effectue en décodant dynamiquement le chemin optimal suivant le critère de maximum de vraisemblance. Les scores obtenus montent la robustesse de l'approche proposée.

1. Introduction

Ce travail de recherche s'intéresse à la reconnaissance automatique de l'écriture amazighe. L'objectif est de concevoir un système de reconnaissance de texte et de document pour des fins de promotion de la langue et de la culture amazighes. Récemment des travaux de recherche basés sur plusieurs approches sont menés dans la littérature pour l'automatisation de la lecture de cette écriture. Une bonne synthèse est donnée dans (Es Saady, 2012). Nous avons initié l'utilisation des HMMs pour la reconnaissance de caractères Tifinaghes dans nos travaux antérieurs (Amrouch *et al*., 2009; Amrouch *et al*., 2010). L'approche préalablement proposée porte des limites et montre des résultats moins encourageants.

Dans ce papier, nous présentons un système basé sur une approche qui exploite les caractéristiques et les spécificités morphologiques de la langue amazighe par une modélisation markovienne optimisée par des algorithmes fondés sur la programmation dynamique (Herman et Ortmanns, 2000; Arica *et al*., 2001). Les caractéristiques exploitées sont extraites à partir des tracés des caractères par une technique de localisation implicite des segments qui le composent. Pour ce faire, nous avons utilisé les points d'intérêts des squelettes. Dans la phase

de classification nous avons déployé le chemin discriminant fondé sur la programmation dynamique opérant au niveau de graphe des segments. Nous présentons par la suite, les graphèmes de base des caractères Tifinaghes déployés par notre système, les composantes du système et l'approche utilisée dans ses différentes phases. Les résultats expérimentaux ainsi que la conclusion et les perspectives de la méthode sont donnés dans les deux derniers paragraphes.

2. Graphèmes de base du caractère amazighe

L'analyse de la morphologie de l'alphabet Tifinaghe révèle certaines particularités intéressantes, en particulier la redondance des segments horizontaux, verticaux et diagonaux dans la majorité des lettres ; ainsi la redondance des formes circulaires qui ne se différencient que par la présence et la position d'un trait ou de point (voir figure 1).

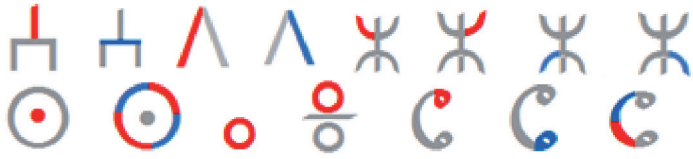


Figure 1 : Les graphèmes fondamentaux observés sur les caractères «O», «D », «Z»,«S», «A», «E» et «C»

On se basant sur ces spécificités, nous avons proposé une liste contenant 10 graphèmes fondamentaux qui constituent les traits de la structure de caractère Tifinaghe (voir tableau 1). Cette liste à été construite à partir de la base de données des patterns amazighes imprimés (Ait Ouguengay, 2008).

S1	S2	S3	S4	S5
S6	S7	S8	S9	S10

Tableau 1: Liste de graphèmes de base

Les caractères Tifinaghes sont théoriquement et visuellement constitués de segments et arcs élémentaires. L'utilisation de ces graphèmes fondamentaux pour décrire la structure d'un caractère Tifinaghe constitue une démarche naturelle. De ce fait, tout caractère de l'alphabet amazighe peut donc être décrit de façon unique et complète par la liste des graphèmes qui le composent. La segmentation directe des caractères imprimés en ces graphèmes constitue

un problème très complexe. Par ailleurs, nous constatons la stabilité de ces graphèmes dans les tracés des caractères Tifinaghes quel que soit le style de l'écriture. En s'appuyant sur ce constat, nous proposons de les considérer directement comme des entités indivisibles de notre modélisation markovienne.

3. Architecture du système développé

Le système de reconnaissance de caractères Tifinaghes que nous avons développé est basé sur une architecture simplifiée illustrée par la figue 3. Cette architecture est de type reconnaissance analytique avec segmentation implicite et apprentissage local au niveau des graphèmes, à travers laquelle le système fonctionne en 2 phases cruciales : l'apprentissage et la reconnaissance. Chacune d'elles inclue un ensemble d'étapes : prétraitement, normalisation, pré-classification et extraction de primitives. En effet, dans la phase d'extraction des primitives, notre système utilise les caractéristiques structurelles extraites des graphèmes de chaque caractère. Ces graphèmes sont exprimés ensuite par un ensemble des modèles. Dans l'apprentissage, nous avons construit un modèle globale ergodique de tous ces graphèmes. Le meilleur alignement à travers ce modèle sera déterminé pendant la classification. Nous détaillons les composantes principales de cette architecture dans les sections (figure 2).

4. Prétraitements

Dans cette section, nous avons effectué une série des prétraitements sur l'image de caractère permettant d'isoler un caractère à partir d'un texte amazighe (voir figure 4). Pratiquement, ils comprennent : la bianarisation, la réduction du bruit, la correction d'inclinaison des lignes de texte, la segmentation d'un texte en caractères, et enfin la normalisation de la taille et la squelettisation.

Pour la binarisation, nous avons opté pour la méthode d'Otsu (Otsu, 1978). Quant à la réduction du bruit, nous avons appliqué le filtre moyenneur (Katkovnik *et al*., 2003). Et comme l'écriture amazighe n'est pas cursive, en plus ne possède pas les ascendants et descendant donc une simple application des projections horizontales et verticales permettent d'isoler le caractère.

La normalisation se résume en 2 étapes. Les images de caractères sont d'abord transformées en images de même taille (48*48) en utilisant un algorithme de normalisation de la taille (Srihari et Keubert, 1997). Par la suite, elles subirent une opération qui consiste à supprimer le vide préexistant entre ses bords et les caractères eux même de telle sorte à avoir des objets recadrés. Pour la squelettisation, nous avons employé l'algorithme de Hilditch (Hilditch, 1969).

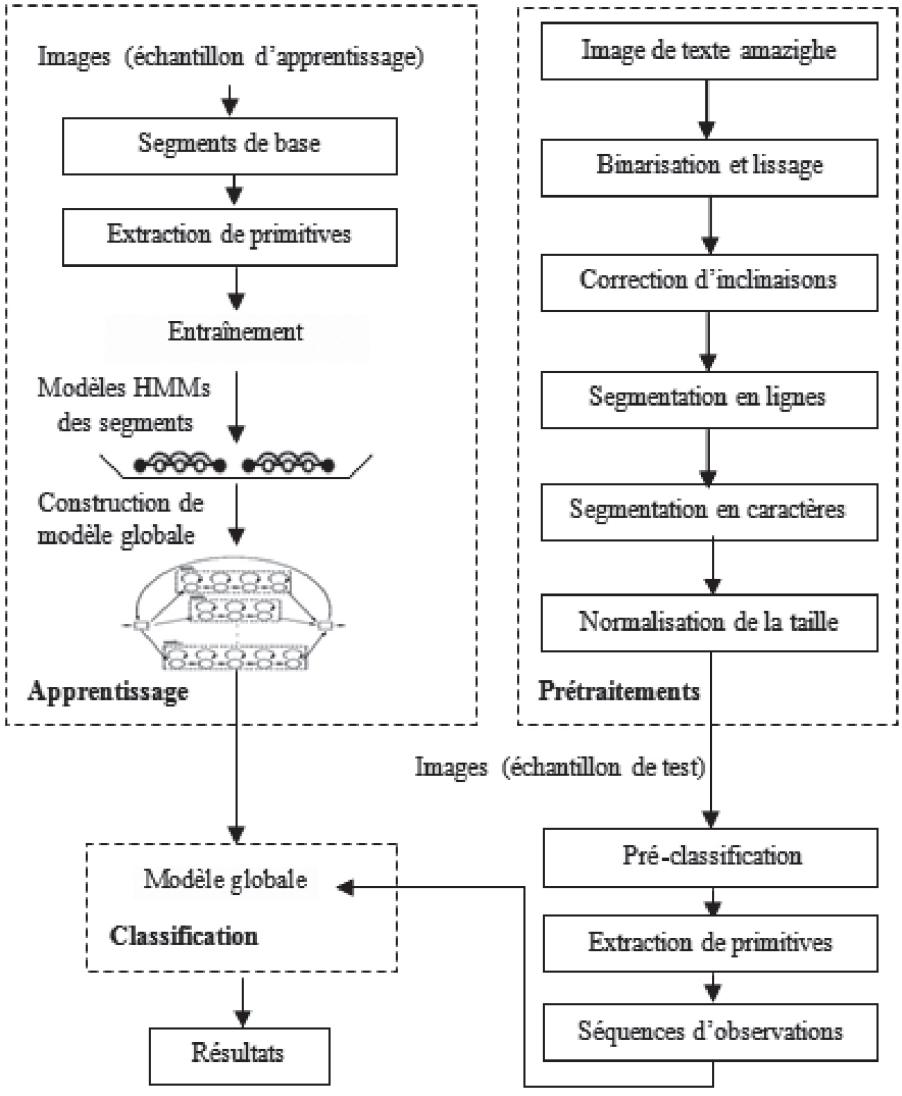


Figure 2 : Synopsis du système développé

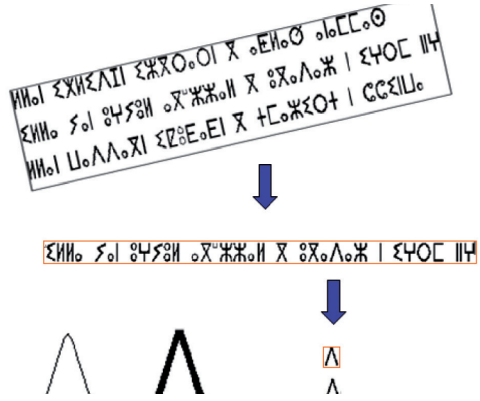


Figure 3 : Localisation et extraction d'un caractère isolé à partir du texte amazighe

5. Pré-classification des caractères

Notre méthode est basée sur les points d'intérêts de squelette. Comme certaines caractères ne les possèdent pas, notamment ceux qui contiennent seulement les formes circulaires, alors l'idée principale est de pré-classer les images de caractères en 2 groupes : formes circulaires (ⴰ, ⵔ, ⵗ, ⵙ, ⵚ, ⵛ) et formes non circulaires (ⵍ, ⵀ, ⴷ, ⵅ, ⵏ, ⵜ). Dans la littérature, il existe plusieurs méthodes qui répondent à cette question. La plus usuelle est la transformation de Hough (Maitre, 1985). Cette méthode semble efficace mais couteuse en terme de temps et de mémoire. De ce fait, nous avons proposé un algorithme sélectif basé sur la combinaison des dérivées secondes et le nombre de point d'intérêt de la courbe de tracé des caractères. Il est donc organisé en niveau successif de décision permettant le filtrage progressif des décisions et réduction de l'ambiguïté en utilisant l'indice de classification obtenu par :

$$w = \frac{\sum_{i=1}^{N-1} \sqrt{Courbe^2}}{N}$$

où N_{pi} : le nombre de point d'intérêt.

Cet algorithme est illustré par la figure 5 ci-dessous:

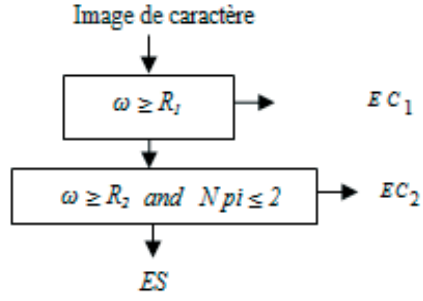


Figure 4 : Diagramme de l'algorithme proposé

Le premier niveau consiste à sélectionner les boucles dotant d'un indice supérieur ou égale au rayon fixe R1=Hauteur/2 qui correspond aux caractères (ⵇ, ⵙ, ⵜ, ⵔ, ⵟ, ⵠ) (voir figure 6 (a)). Cependant dans le deuxième niveau, nous avons fixé de la même manière un autre rayon R2=Hauteur/4 pour les caractères (ⵄ, ⵅ, ⵛ, ⵃ), mais l'utilisation du paramètre ω uniquement conduit à des confusions entre les formes (ⵄ, ⵅ, ⵛ, ⵃ, ⵣ, ⵠ, ⵅ, ⵠⵯ, ⵅⵯ) (voir figure 6 (b)). Ceci appuie l'idée de la localisation des boucles sur laquelle est fondé notre algorithme. Afin de résoudre cette ambiguïté, nous avons proposé de combiner dans ce stade l'indice ω et le nombre N_{pi} total des points d'intérêt de la forme pour filtrer les formes qui possèdent plus que points d'intérêts (voir figure 6 (c)). A l'issue de cette technique, nous avons obtenu deux classes de caractères :

EC={EC_1= (ⵇ, ⵇ, ⵝ, ⵔ, ⵟ, ⵠ), EC2= (ⵄ, ⵅ, ⵛ, ⵃ)};

ES={ⵡ, ⵜ, ⵚ, ⵣ, ⵀ, ⵏ, ⵍ, ⵅ, ⵉ, ⵔ, ⵞ, ⵟ, ⵡ, ⵟ, ⵟ, ⵠ, ⵔ}.

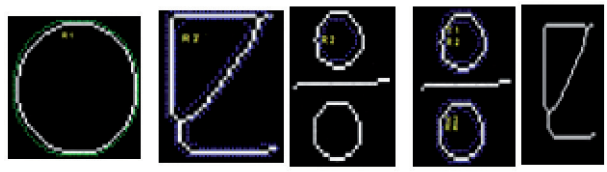


Figure 5: Résultat obtenus par notre algorithme

Nous avons poursuivi ce processus de filtrage pour le groupe EC dans le but de partitionner les symboles selon le nombre N_{cc} de composantes connexes qui le composent. La démarche est résumée par le schéma ci-dessous :

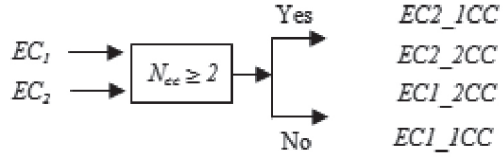


Figure 6: Algorithme de composantes connexes

Finalement, nous avons obtenu les groupes des caractères suivants :

EC1_1CC={ⵇ, ⵙ, ⵜ, ⵔ, ⵠ}; EC1_2CC={ⵟ}; EC2_1CC= {ⵄ,ⵛ} ; EC2_2CC {ⵅ, ⵃ} et ES.

Par la suite, chaque classe a subi un traitement approprié à sa morphologie avant d'être soumise au module commun d'extraction. Les traitements qu'ont été effectués sont : l'extraction des composantes connexes pour les caractères de la classe EC1_2CC et ceux de EC2_2CC et une décomposition de l'image du caractère en 4 parties par leurs axes centrales verticale et horizontale pour les groupes EC1_1CC et EC1_2CC.

6. Extraction de primitives

Dans notre système, nous avons opté pour les caractéristiques structurelles qui décrivent les propriétés topologiques et géométriques de l'écriture, souvent extraites directement à partir de squelette de tracé et qui représentent globalement : le nombre d'Euler ; la longueur minimale et maximale des graphèmes ; les surfaces, les diamètres et les périmètres ; les segments de droite et leurs attributs (position, centre de masse, orientation, ...) ; des arcs, boucles, et concavités ; mesures des courbures et orientations principales ; mesure d'excentricités, solidités et étendues ; points des jonctions, angularités, et terminaux ; le premier et le deuxième moment de Hu (Hu, 1961).

Afin de pouvoir compter certaines de ces paramètres, nous avons procédé à analyser le squelette de chaque caractère pour la localisation de ses points d'intérêts (points d'extrémités, les points d'intersection, les points d'inflexion) en se basant sur le nombre de transition d'un pixel noire à un pixel blanc (0 →1) sur son *8*-voisinage *P* (Arrivaul, 2008).

Une fois les points d'intérêts sont déterminés sur le squelette de caractère, nous utilisons un algorithme du suivi de contour afin d'extraire les séquences d'observation de chaque caractère.

Un point d'intérêt x_i est définie par ses cordonnées (x_i y_i). Chaque segment est représenté par une suite de pixels de squelette délimité par deux points (x_{i1} $y_{i\,1}$) et (x_{i2} y_{i2}).

Au cours du suivi de contour expliqué précédemment, nous calculons un ensemble d'indices structurels sur chacun des différents segments qui composent le caractère en utilisant les boîtes englobantes de hauteur $h=|y_{i2}-y_{i1}|$ et de largeur $l=|x_{i2}-x_{i1}|$. Cependant, dans le cas des segments verticaux où la hauteur est nulle ($h=0$) et les segments horizontaux où la largeur est nulle ($l=0$), nous avons considéré des fenêtres avec des dimensions $h=10$ et $l=10$ respectivement (voir figure 9).

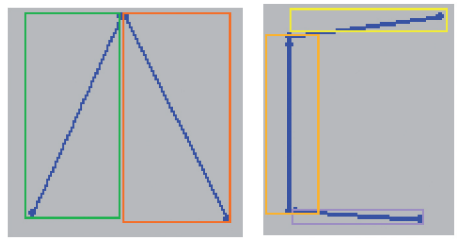


Figure 7 : Exemples de boîtes englobantes

En fin, pour chaque caractère, nous obtenons un vecteur de primitives qui regroupe les séquences générées à partir de chacun des segments qui le composent.

7. Modèles HMMs de segments

L'approche proposée est analytique, basée sur la modélisation des segments de la liste des graphèmes présentée préalablement par des modèles de Markov cachés (Rabiner, 1989) ce qui donne en totalité 10 modèles. Le modèle $\lambda = \{N = 3, M, \Pi, A, B\}$ d'un segment est du type gauche droite (voir figure 10).

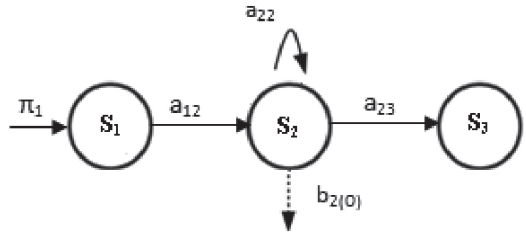


Figure 8 : Le modèle HMM gauche droite des segments

7.1. Apprentissage

Pendant l'apprentissage, nous avons utilisé la procédure classique de Baum-Welch (Augustin, 2001), qui permet de réestimer le HMM de chaque segment λs jusqu'à ce que la vraisemblance de générer la séquence d'observations soit maximale. Le meilleur HMM trouvé est enregistré. Par la suite, l'ensemble des modèles obtenus sont concaténés pour former le modèle globale ergodique de notre système.

7.2. Classification

A l'issue de l'étape d'apprentissage, nous avons obtenu un modèle globale ergodique modélisant l'ensemble des graphèmes de la liste proposée. Chaque chemin dans ce modèle représente une séquence de segments. Par conséquent, la reconnaissance d'un caractère se fait par la recherche de meilleur chemin dans ce treillis des segments. Nous avons employé l'algorithme Viterbi qui est fondé sur la programmation dynamique. Il consiste à décoder la meilleure séquence d'états candidates en se basant sur un critère de maximum de vraisemblance.

8. Expériences et résultats

Pour valider le système proposé, nous avons effectué des expérimentations significatives sur la base de données de patterns de la graphie amazighe (contient 19437 caractères multi fonts c-à-d 627 échantillons x 31 classes). Nous avons définies à partir de cette base deux ensembles distincts de données, un ensemble A (A=2/3) pour l'apprentissage et un ensemble B (B=1/3) pour les tests.

Plusieurs tests ont été effectués, pour évaluer le taux de reconnaissance du système en fonction de : nombre d'états et de nombre de mélange de gaussiennes. Le tableau 2 présente les résultats obtenus de ces tests sur cette base.

Nombre d'états	*3*	*5*
Nombre de mélange de gaussiennes	*1-2-3*	*1-2-3*
Taux de reconnaissance	*98,21%*	*98,52%*

Tableau 2 : Taux de reconnaissance

Ces résultats montrent un taux d'erreur de 1,48% avec un modèle de topologie de 5 états. Nous estimons que les erreurs de reconnaissance sont attribuées, d'une part, aux méthodes utilisées pour la pré-classification et à la détection des points d'intérêts, et d'autre part, à la déformation de certains caractères dans certaines fontes et à l'insuffisance des caractéristiques utilisées pour mieux décrire chaque segment. En vue de remédier à ces problèmes pour diminuer le taux d'erreur de notre système et par conséquent avoir un système fiable, nous avons proposé, une autre variante de ce système, qui consiste à adopter la transformation de Hough pour séparer les lettres circulaires et les non circulaires. En plus, cette version s'appuie sur une autre technique pour déterminer les points d'intérêts de la forme, la technique utilise la notion de déviation maximale au niveau de la courbe du caractère.

De la même façon, pour évaluer la version améliorée du notre système, on recommence la première expérience précédente. Le Tableau 3 montre les résultats obtenus par la version améliorée sur la base de patterns de la graphie amazighe.

Nombre d'états	*3*	*5*
Nombre de gaussiennes	*1-2-3*	*1-2-3*
Taux de reconnaissance	*98,41%*	*98,76%*

Tableau 3: Taux de reconnaissance

Dans ce cas, on constate que le taux est de 98, 41% par une topologie de type gauche droite à 3 états et le même pour tous les modèles gaussiens utilisés (1,2 et 3). Soit un gain de 0.20% sur le score avec la nouvelle version. En outre ce taux s'élève à 98,76 par la topologie du type gauche droite à 5 états et le même pour tous les modèles gaussiens utilisés (1,2 et 3). C-à-d cette fois-ci un gain de 0,24. En comparant ces taux de reconnaissance obtenus par la version améliorée de notre système avec ceux obtenus par la première version, nous soulignons une amélioration qui s'échelonne entre 0.2% et 0.25%. De ce fait, les techniques déployées lors des phases pré-classification et extraction de primitives, notamment celle qui concerne la localisation des segments de caractères, ont une influence significative sur la performance du système. Par ailleurs, cela démontre que l'emploie de la technique de déviation maximale au lieu de la méthode classique pour détecter les points d'intérêts permet une augmentation au niveau des scores obtenus.

9. Conclusion

Dans ce papier, nous avons proposé une solution au problème de la reconnaissance automatique de caractères Tifinaghes imprimés dans un vocabulaire limité de segments. La solution apportée est fondée sur les caractéristiques morphologiques du caractère Tifnaghe et la programmation dynamique par des HMMs continus. Les résultats obtenus sont tout à fait encourageants. Ils montrent que les HMMs continus sont plus robustes. Cependant les inconvénients de cette approche résident dans : (1) la détection des points d'intérêts pendant l'extraction des caractéristiques semble contraignante pour certains styles d'écriture utilisés ; (2) le choix des éléments de vocabulaire, qui ont été établis de façon à couvrir l'essentiel des formes de segments rencontrés dans l'alphabet Tifinaghe, impose une contrainte sur les styles admissibles par le système. En effet, un caractère dont un de ses segments ne correspondrait à aucun élément de ce vocabulaire ne pourra pas être reconnu par le système. Pour remédier à ces problèmes, l'intégration d'autres caractéristiques, augmentation de taille de vocabulaire, la proposition d'une autre méthode de segmentation des caractères en graphèmes et l'utilisation des HMMs en combinaison avec d'autres classifieurs notamment le MLP peuvent constituer une perspective que nous envisageons pour l'amélioration des performances de notre système.

Références

A. Ait Ouguengay (2008). Elaboration d'un réseau de neurones artificiel pour la reconnaissance optique de la graphie amazighe, Phase d'apprentissage. *5ème conférence sur les systèmes intelligents : Théories et applications* (SITA'08), Rabat, Maroc.

M. Amrouch, Y. Es saady, A. Rachidi, M. Elyassa, D. Mammass (2009). Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform. *ICMCS'09*, Ouarzazate, Maroc.

M. Amrouch, M. Elyassa, A. Rachidi, D. Mammass (2010). Handwritten Amazigh Character Recognition Based On Hidden Markov Models. Accepté dans *International Journal on Graphics, Vision and Image Processing*.

N. Arica, T. Fatos, V. Yarman (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man and Cybernetics - part C: Applications and Reviews*, 31(2): 216-233.

D. Arrivaul (2008). Apport des Graphes dans la Reconnaissance Non-Contrainte de Caractères Manuscrits Anciens. Rapport de thèse, Traitement du Signal et des Images, Université de Poitiers.

E. Augustin (2001). Reconnaissance de mots manuscrits par systèmes hybrides Réseaux de Neurones et Modèles de Markov Cachés. Thèse, Paris.

Y. Es Saady (2012). Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents amazighes. Thèse de doctorat, Université Ibnou zohr, Agadir, Maroc.

N. Herman, S. Ortmanns (2000). Progress in dynamic programming search for LVCSR. *IEEE*, 88(8): 1224-1240.

J. Hilditch (1969). Linear skeletons from square cupboards. *Machine Intelligence 4, B. Meltzer, D. Michie, Eds*., pp. 404-420.

M.K. Hu (1961). Pattern recognition by moment invariants. *IRE Transactions on Information Theory,* vol. 8, issue: 2, pp. 179-187.

V. Katkovnik, K. Egiazarian, J. Astola (2003). Application of the ICI principle to window size adaptive median filtering. *Signal Processing*, n° 83, pp. 251-257.

H. Maitre (1985).Un panorama de la transformation de Hough, École Nationale Supérieure des Télécommunications, Laboratoire Image, Département Images, Sons et Vidéo, traitement de signal, vol. 2, n° 4.

N. Otsu (1978). A threshold selection method from grey-level histograms. *IEEE Trans. Syst. Man. Cybern*, vol.SMC-8.

L. Rabiner (1989) A Tutorial on Hidden Markov Modelsand Selected Applications in Speech Recognition. *IEEE*, 77(2): 257-286.

S.N. Srihari, E.J. Keubert (1997). Integration of handwritten address interpretation technology into the united states postal service remote computer reader system. *ICDAR*, pp. 892-896.

La Conférence Internationale sur les Technologies d'Information et de Communication pour l'AMazighe (TICAM) est un événement bisannuel organisé à l'Institut Royal de la Culture Amazighe (IRCAM), par le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC).

Instauré depuis 2004, TICAM est le rendez-vous des scientifiques, chercheurs et professionnels qui oeuvrent dans le domaine de la technologie de l'information et de la communication appliquée aux langues naturelles et particulièrement à la langue amazighe.

La conférence a pour objectif la valorisation des travaux de chercheurs nationaux et internationaux, mais aussi la promotion des travaux de jeunes chercheurs, tout en offrant un panorama représentatif de l'état de l'art comme des perspectives les plus motivantes pour le développement de ce domaine.

Cette 5ème édition a été également une opportunité pour faire intervenir des conférenciers invités à exposer les avancées dans leurs domaines, sur les plans théorique, applicatif et expérimental, et pour organiser des ateliers de démonstration pour mettre en lumière les défis de la complexité du développement des langues naturelles et de donner la parole aux chercheurs lors des sessions orales et des présentations de posters.

40 Dh